Data Management Plan
Project: The influence of plant functional types on ecosystem responses to altered rainfall
PI: Elsa Cleland (UCSD), Co-PI: David Lipson (SDSU), Co-PI: John Kim (SDSU)

Training
At the start of the funding period the PIs, senior personnel, technician and students on the project will convene a dedicated data management meeting. At this time the PIs will set out naming, processing and storage conventions for all data collected at the experimental and observational sites, as well as conduct training in annotating datasets with necessary metadata. All participants will be trained in data management best practices (*e.g.* Borer *et al.* 2009). This training will be reiterated at a yearly data management and analysis meeting, to remind participants of the conventions and train any new participants.

Collection
Datasets to be collected in the experimental and observational components of the research are described in the Project Description. Most datasets will be collected 1-3 times per year (i.e. production and decomposition, ecophysiological functional traits, soil extractable nutrients and mineralization rates, soil microbial community composition and function). Temperature, light availability and soil moisture at multiple depths in the experiment will be logged every 15 minutes, these data will be stored on local data loggers and downloaded every two weeks.

Processing
Data originally recorded on physical paper datasheets will be transferred each day into spreadsheets using non-proprietary software (*e.g.* open office platforms stored as ASCI files, .txt or .csv formats). Data will immediately be checked for outliers in the R statistical program, and any outliers will be checked against the paper copies for transcription errors. Paper copies will be kept on file for at least 10 years. All field-collected data will be stored on the Division of Biological Sciences Twinlake server which is backed up daily and is duplicated on servers in two separate buildings on campus. Data collected by project staff based at SDSU will be stored on the Field Stations Program Ringtail data server, which features an expandable RAID system configured to store data with redundancy across multiple drives. Furthermore, for added redundancy, coordination, and availability all project data will be shared (copied) periodically between UCSD and SDSU, in a manner consistent with university policy.

Analysis
Data analysis of field collected data will be scripted and extensively annotated in the R statistical program and the scripts will be stored on the server along with the datafiles. DGVM simulation runs will be performed on a high performance parallel computing platform, a 96-node Linux cluster, maintained jointly by USFS Pacific Northwest Research Station and Oregon State University. Simulation output is in NetCDF format, a data format popular in climate research, readable with many free software programs. DGVM output will be analyzed and displayed with the ESRI ArcGIS software suite.

Documentation
All datasets will be annotated with meta-data. As data are generated they will be entered into Morpho, a free resource for associating Ecological Metadata Language (EML) with archived

Data Management Plan

The data coupled with the research project will be systematically managed. The team has multiple backup servers to protect our research findings, and publicly available internet resources to share our results. All aspects of the research will be carefully tracked, stored, and published. The work detailed in the preceding proposal can be anticipated to produce three broad categories of data: computer software, subjective test data, and models. The computer software category consists of video encoders and decoders, including error concealment and forward error correction algorithms; and channel-aware encoding and rendering adaptation algorithms. The subjective test category includes data from the human observer experiments in the packet loss visibility study, as well as data from the quality of experience in channel-aware rendering video study. The model category includes the packet loss visibility model, the cross-layer model with mobility considerations and the quality-of-experience model.

The algorithm development progression will be logged through both handwritten research notebooks as well as digitally generated documents. To ensure the safety of the data, we will use the Video Processing group's existing file server to periodically backup the materials. A Structured Query Language (SQL) database will be created to track the digital documents. We plan to package our algorithm in a MATLAB toolbox, as well as a self-contained software complete with a user interface. All of the computer code produced during the project will be written using the latest version of MATLAB, as well as C/C++, when appropriate. Codes will be developed using volume shadow copy technology, which will allow the recovery of prior iterations for quality control.

The results of the research performed under this proposal will be disseminated primarily through publication in research journals and conference presentations. All of the computer software, subjective test data and model parameters will be available to interested parties upon request, and will be transmitted electronically via e-mail.

All electronic data generated by proposal research will be redundantly archived. Locally, the laboratories have secure servers on which all information is stored. The server hard drives are set up in a RAID that is capable of full recovery even in the case of multiple simultaneous disk failures. Additionally, the server drives are backed up by servers operated by School of Engineering IT. This will allow full recovery of data in the event of catastrophic failure of the local laboratory servers. All of these systems will be in place for the 3 year minimum prescribed by the NSF, as well as the foreseeable future following that.

datasets. It will be the responsibility of each researcher to annotate their data with metadata, and it will be the responsibility of the PIs to check weekly (during the field season, monthly otherwise) with all participants to assure data is being properly processed, documented, and stored.

Data Products, Curation and Data-Use Policy

All raw data will be made freely available by the time of publication or the end of the funding period, consistent with NSF policy. When data are associated with a publication, the raw data and associated analysis R scripts will be archived in a source such as Ecological Archives or Dryad. DGVM outputs from the project will be archived in a recognized national data center, such as the ORNL DAAC, ESRI DataBasin, or UNH EOS-Webster, making data freely available for download. At the end of the funding period all datasets will be entered together as a data package into the KNB database (Knowledge Network for Biodiversity) and/or the DataOne database, with the goal of archiving the data together for perpetuity. Prior to the end of the funding period data will be made available by request with the stipulation that if the data are used in publication then the researchers that collected the data need to be informed of the planned use and be offered authorship as appropriate.

Past activities

All senior personnel have been active in efforts to efficiently document and synthesize ecological data. As a postdoctoral researcher at the National Center for Ecological Analysis and Synthesis (NCEAS) PI Cleland participated in the SEEK-BEAM working group (Science Environment for Ecological Knowledge - Biodiversity and Ecological Modeling and Analysis), an ecoinformatics effort to test the implementation of Morpho and other ecoinformatics tools developed at NCEAS. She also led an NCEAS Distributed Graduate Seminar to teach data management and synthesis techniques in the field of Functional Ecology. Collaborator John Kim has worked as data manager for SDSU Field Stations Program and for SDSU Global Change Research Group. He has also taught data management workshops sponsored by NSF: for postdocs and new faculty through SEEK, and for field stations staff at RCN Resource Discovery Initiative for Field Stations Training. Finally, both PIs have a demonstrated commitment to publishing data and associated meta-data. PI Cleland published all data associated with her synthetic work at NCEAS (Cleland et al. 2008), and CoPI Lipson submitted all data from a recently completed NSF grant (OPP 0421588) to the ARCSS site at EOL (http://www.eol.ucar.edu/projects/arcss/).

Note: For references please see main references document

Data Management Plan

Types of data

1. Raw data: Nearly all experiments under the proposal involve electrophysiological recordings in behaving rats. The primary (i.e., raw) forms of data are: 1) waveform recordings in the format of the .plx files (Plexon MAT recording system); 2) video and LED tracking files in the form of .avt and .dvt files (the latter is output of the animal's head position in x,y coordinates collected at 60 Hz by the Plexon Cineplex Studio program); 3) histological data in the form of Nissl-stained brain slices (to localize recording or microinjection sites). The applicant's laboratory houses a microscope with wide-field lenses and a digital camera and regularly practices archiving of relevant digital photographs of brain slices..

2. Pre-processed data: Nearly all experiments will demand categorization of spike waveform data into subsets corresponding to single neurons. In the field, the method and outcome of such categorization is often a matter of interest. Waveform categorization output takes the form of .plx files (Plexon Inc.). Nearly all experiments involving navigation and tracking of the animal are put through a behavioral 'screening' procedure whereby portions of the behavioral record are selected for further analysis (e.g., identification of uninterrupted runs along a track). This process is carried out in Matlab and the interpretation of the output .mat files requires only a key to the behavioral coding scheme (easily available via the Nitz laboratory).

3. Analyzed data: All electrophysiological and behavioral data is analyzed using the commercially available program Matlab. The resulting .mat files involve analyses of many types. However, nearly all experiments involve creation of a 'ratemapping' analysis wherein the firing rates of neurons according to positions along a path are determined.

Data and Metadata standards

Electrophysiological data take the form of .plx files (Plexon Inc. standard), .avt files (overhead videos of recording arenas), .dvt files (Plexon Inc. standard for position-tracking data), or .jpg files (digital photographs of stained histological brain slices). Pre-processed data (see section above for explanation) comes in the form of .plx files (Plexon Inc. standard for categorized spike waveforms) and .mat files (Matlab standard – output of screening/filtering of tracking data). Analyzed data (see above section for explanation) takes the form of Matlab .mat files. Metadata comes primarily in the form of surgery and recording logs where the nature of the particular experiment (stereotaxic targets, channel configurations, behavioral apparati utilized, recording and animal nos., etc.) is written and drawn on paper. Such recording logs are copied once a month and stored offsite. Access could be made via digital photography if requested. Other metadata include the specific custom-written analysis programs utilized and information concerning amplifier settings and channel configurations. The latter are contained within the .plx electrophysiological data files. Custom-written analysis programs (Matlab .m files) are regularly copied and saved to several different on-site and off-site computers.

Policies for access and sharing

All data, metadata, and analyses collected under the proposed experiments will be made publicly available as per NSF guidelines within 2 years of collection via published manuscripts, publicly available final reports to NSF, and/or from data archives at UCSD's Department of Cognitive Science. The PI will take guidance from NSF concerning the right to use the data prior to opening it up to wider use. There are no ethical or privacy issues involved in sharing of this type of data and it is unlikely that sharing will incur more than modest cost. The most likely data-sharing scenario concerns the categorization of spike waveforms as this process does not enjoy agreement in standard over the entire field. Possible requests

may also come from neurophysiologists desiring to double-check the validity of a published analysis using categorized spike waveform data and behavioral tracking data.

Each of the formats by which data and metadata are archived can be used with relative ease by anyone requesting access. Plexon .plx files can be converted for use in standard programming environments such as Matlab through use of software downloaded from the Plexon website. Tracking data comes in the form of .dvt files which can be directly loaded into Matlab or Notepad. Video and digital photographic files are in formats (.avt, .jpg) that can be opened using a large number of freely available programs.

Data archiving

All electronic data are saved in triplicate with two copies kept on-site (UCSD Cognitive Science Department) and another kept off-site. As described, all data, whether in electronic or paper form, are copied, organized by animal number and recording day, and archived both on-site and off-site. There is no plan to destroy any collected data as the archive is not burdensome in cost or space. As such, data archives can be expected to be available for at least a period of ten years subsequent to publication of the relevant findings. As described above, the PI's attention to detail in archiving and the ease with which data can be shared in usable format makes data-sharing a relatively simple process for the proposed experiments. In the event that a simple email or snailmail delivery of electronic files cannot be accomplished, the PI has access to the department's system administrator who can set up an FTP server. A third avenue for sharing is via the SRB (Storage Resource Broker)/iRODS (integrated rule-based data) system developed at the San Diego Computer Center for sharing of data from NSF projects such as the Temporal Dynamics of Learning Center.

Data Management Plan, PI Daniel Rogalski

It is noted in the NSF's document *Division of Mathematical Sciences (DMS) Advice to PIs on Data Management Plans* that "For many proposals to DMS, a statement that no data management plan is necessary will suffice, provided that a clear justification for this claim is given."

As suggested, we feel that this proposal requires no data management plan, since the research involves little data external to the mathematical results themselves, which will be disseminated through publication as well as through posting to the arXiv.

We do note that as part of the previously funded project, we wrote some short computer programs in Maple as a tool for one of the research papers we published joint with J. Zhang. It is possible that we may continue to produce some computer code to aid in one or more of the research questions in the current project proposal, especially the questions related to Calabi-Yau algebras. We made the code we produced in the past freely available on the PI's website. We plan to similarly share openly any further code produced which we deem to be of possible interest or use to others.

Data Management Plan

1. Data and Metadata

1a. Types of data collected during this project

Field-based data

1. Time-series observations of soil moisture, O_2, temperature, electrical conductivity, redox potential, and water table level via data loggers and cellular connection to web server.
2. Automated soil surface flux and soil subsurface CO_2 concentrations via data loggers.
3. Meteorological data from on-site meteorological station via cellular connection to web server.
4. Soil and GHG samples for laboratory analysis.
5. ECa data collected during periodic campaigns.

Research team members will perform QA/QC on data above when they are downloaded from the data loggers and/or transferred from field books to digital files.

Laboratory Data

1. Soil water chemistry, including nitrate and ammonium.
2. Greenhouse gases including N_2O, CO_2 and CH_4.
3. Redox buffer capacity assays.

In addition to the analytical results, data will be collected to verify proper sampling protocol, transportation and storage, as well as QA/QC of analytical data. Routine analytical procedures are formally written with respect to the specific details of the PI's laboratory and are stored in the laboratory, both in digital and printed format as a reference for all group members. Our lab has developed a quality assurance protocol that involves routine collection and analysis of blanks and duplicates; routine analysis of analytical blanks and duplicates, and cross calibrated with certified standards. Method detection limits and recovery efficiency are also determined where appropriate. This quality assurance program will be adapted and applied to the project proposed herein. Although this intensive quality assurance protocol requires an investment of time and resources, it results in a high level of confidence in the analytical data. It also serves as an important training tool for the post-doc and undergraduate students. Most analytical equipment is connected to a computer for data acquisition. All computer generated experimental data will be collected in associated computer files stored on the computer associated with the instrument as well as backed up on Dropbox.

Existing Data

Freely available data (e.g., digital elevation maps, aerial imagery, climate projections, and local meteorological) will be accessed and used. In these cases, we will note the data source and access date in our metadata and store the data themselves with our project files.

1b. Data and metadata formats and standards

All data described above will include a plain text file containing relevant metadata using the Ecological Metadata Language standard. This metadata will include date and location of data collection, collection personnel, collection methods (instrument type, calibration date(s)), description of the ongoing experiment or ambient conditions during collection, a description of the data included, an indication of the use of *NaN* ("Not a Number") to represent a missing observation, any QA/QC performed on the data and the results of those analyses, quantification of error in measurements (if known), and contact information for each of the PIs. Morpho (https://knb.ecoinformatics.org/#tools/morpho) will be used to build links among data and metadata for easy sharing and accessibility.

A local datum and coordinate system will be established using a relatively permanent benchmark at the site (e.g., a spike driven into a large tree) and all measurement and modeled locations defined relative to this point. We will include a description of the local datum and coordinate system in the metadata file for all data. The local system will be integrated into global coordinate system for data sharing and publication of results.

Field data, laboratory data, existing data used for the project, and model results will be output from the models and stored in ASCII format. For commercial software files, a plain text file documenting

the software (title, version) used will be stored with the results and the proprietary model files, if any. For open-source ("freeware") and any custom Matlab code for this project, a copy of the software or code will be archived with the data. Original field and laboratory notes will be photographed or scanned weekly and archived alongside digital copies of the data. Field data sheets will be scanned weekly and field data will be convert to digit files and stored as ASCII.

2. Storage and preservation resources and facilities

Prior to long-term archival, data will be stored on a servers owned by the University of Nebraska-Lincoln including back-up to a second, off-site location. Backup copies of all data will also be stored on Dropbox. Data will be password protected on university servers and Dropbox with access provided to the project team for upload. A copy of all field data collected will also be stored on a dedicated hard disk in [the lead-PI's] office. Original field and lab notebooks will be secured in locked cabinets in University offices. Data will be synchronized between storage locations on a weekly basis, at minimum.

At the conclusion of the grant, long-term preservation of and access to project data will be provided by the University of Nebraska-Lincoln Data Repository (dataregistry.unl.edu). Briefly, the UNL Data Repository is hosted by UNL Libraries and provides a permanent URL (via DOI), data integrity checks, secure and replicated storage (multiple copies of data stored onsite and at a remote server location), accurate metadata, and global accessibility. All data and metadata described in Section 1 of the Data Management Plan will be archived at this site.

3. Data and metadata dissemination methods

During the project and prior to publication of significant results (or prior to public release), access to the data will be granted by agreement of all PIs, with a request to the PI made at their office phone number or their university email. Requests will be considered on a case-by-case basis. Data will be provided via email or web-based download to grant these requests. There will be no charge for data access.

Following its public release, data will be discoverable and accessible through the UNL Data Repository, and metadata will be shared with indexing services to promote discoverability. Links to supporting datasets will be included in any related publications.

4. Data sharing and access policies

Data will be made available immediately after publication of significant results, and no later than 18 months after the completion of the project. Data will be preserved for a minimum of 3 years after it is made available publicly.

No IRB, ethical, or personal privacy issues exist for the data collected in the proposed research plan. Land owners will remain anonymous in any reports or publications; only approximate geographic locations will be given on any published maps.

5. Data management roles and responsibilities

Data will be deposited in the UNL Data Repository—with the necessary embargoed time limits (maximum of 18 months)— by [the lead-PI] at the conclusion of the grant. [The co-PI] will assume duties for data archiving if [the lead-PI] leaves the project before its conclusion.