Data Management Plan

I. Types of Data Produced

This project will produce many diverse datasets. The Climate team will work with current and archived climate data from weather stations throughout the state, including 5-minute, hourly and daily conditions for air temperature, relative humidity, wind direction and speed, soil temperature at 2-inch depth, solar radiation, and rainfall. Microclimate data will be collected by Doppler radar. Climate models will also use data from the North American Regional Climate Change Assessment Program, Missouri Mesonet, the PRISM grid (parameter-elevation regressions on independent slopes model), the National Elevation Dataset 30m Digital Elevation Models grid, the Pennsylvania State University Soil Information for Environmental Modeling Ecosystem Management database, and the National Land Cover Dataset 2001. The climate team will also collect soil redox potential, soil moisture, and pH using *in situ* probes.

The plant team will collect a wide variety of data. Genomic data will be generated at a core sequencing facility. The LemnaTec Scanalyzer 3D platform will be used to collect and analyze images of plants in controlled environments, including RGB (human visible), NIR (near infrared) and fluorescent, followed by digital construction of 3D images. RGB imaging allows quantification of plant color and morphological characteristics, such as leaf area, stem diameter and plant height. NIR imaging enables visualization of water distribution. Fluorescent imaging is used to quantify chlorophyll. High throughput plant phenotyping data will be collected in the field using autonomous ground vehicles (AGV) equipped with cameras for visible light, fluorescence, multispectral, near infrared, and infrared images; robots for acquisition of leaf tissue for biochemical and/or molecular analysis; sensors for acquisition of gas exchange measurements for analysis of photosynthesis and respiration; and sensors for temperature, humidity, light intensity, wind speed and wind direction. Plant phenotypes to be modeled from the AGV imaging data include color, texture, plant and ear height, leaf and branch angles, leaf width and length, number of nodes, number of ears, tassel branch number, number of pods per inflorescence, size of pods, pubescence density, and venation pattern. Unmanned aerial systems will collect field crop canopy characteristics and individual plant characteristics using several imaging technologies including Color-Infrared, Normalized Difference Vegetation Index, and thermal imaging. Multispectral imaging spectroscopy will be used to collect optical signatures of plants. Webcams will collect images to estimate timing of landscape phenological changes. Additional data includes sensor-generated data (soil moisture, sap-flow, canopy temperature data), real-time water isotope data from transpiring plants from a Picarro isotope analyzer, and model outputs.

II. Data and metadata standards

We will use accepted data format and metadata standards whenever possible. Climate data will be stored in CF-compliant NetCDF format. Doppler radar data will be available as Nexrad level III and IRIS, which can be converted to Universal Format. Genomic data formats include Fastq, Fasta, GFF3, BAM/SAM, VCF. File formats include text/ASCII, standard imaging (e.g. jpg, pgm, ppm, tiff for 2D, ply, blend, mesh, pcd for 3D), imaging for GIS (GeoTiff), video (e.g. mp4, MPEG, avi), binary MatLab (mat). For some data types, metadata content is embedded in data files. For example, webcam image data will be stored in JPEG format, because it includes the ability to store meta-data as EXIF-tags within the jpg format itself, including time-stamps, GPS location, exposure, focal length, focus distance, and what color-correction algorithm has been applied. Similarly, GeoTiff is a public domain metadata standard that allows georeferencing information to be embedded within a TIFF file.

Metadata content and format standards, ontologies and controlled vocabularies for some data types are well-established, while for other data types standards are not widely available. Significant effort of the CI team will be to identify existing data standards, to develop new standards and ontologies, and to assist investigators in applying standards. The Federal Geographic Data Committee (http://www.fgdc.gov/) provides resources for geospatial metadata standards, including remote sensing and meteorology/climatology data. Existing biological ontologies are available through the National Center for Biomedical Ontologies (http://bioportal.bioontology.org/ontologies) and Open Biological and Biomedical Ontologies (http://www.obofoundry.org). The Marine MetaData Interoperability Ontology Registry and Ontology (http://mmisw.org/orr/#b) provides access to ontologies that may be suitable for environmental sensor data. Whenever possible we will cooperate with others who are developing metadata standards for image-based plant phenotyping, which is an emerging area of research. For example, the Phenomics Ontology Driven Data repository (PODD), supports the Australian Plant Phenomics Facilities

(http://www.plantphenomics.org.au/projects/podd/). Other plant ontology resources include the Plant Working Group of the Phenotype RCN (http://www.phenotypercn.org/) and the Plant Ontology Consortium (http://www.plantontology.org/). We will also leverage image analysis resources at iPlant (http://www.iplantcollaborative.org/discover/image-analysis-bisque). Metadata for plant phenotyping will be saved in text files, and will include date, time, location, plant ids (genotypes, individual barcodes) and environmental variables. Metadata for derived data will include information needed to repeat the computations, and related computer code will be fully annotated and stored along with the data sets. We will automate creation of metadata files through features already existing in devices or code developed by the CI Team.

III. Policies for access and sharing

All data will become freely available after publication. Data that is not immediately freely available upon collection will be provided upon request at any time via an online request form that will require users to enter their name, affiliation, description of how they intend to use the data, and to agree not to publish the data without consent. Data will be made accessible to the public via proposed EPSCoR servers (e.g. High Resolution Climate Data), servers in investigator's labs (e.g. the AMOS server at WUSTL and the CCBL server at Danforth), and public repositories (e.g. NCBI Sequence Read Archive and iPlant). We will develop web interfaces for easy identification and access to datasets, including direct download from EPScoR servers and links to datasets in repositories.

Our default policy for newly developed algorithms and software will be to release as free open-source software, so as to maximize the dissemination and impact of the tools. Code and software will be made available in repositories, such as GitHub (https://github.com) and Sourceforge (http://sourceforge.net).

We will seek IRB approval for survey participants in the Community research project. Evaluation is exempt from IRB at MU. With respect to webcam use by the Plant team in publicly accessible locations, we will exclude images that include people close enough to be recognizable. Webcams imaging at field sites will not be publicly accessible, and we will seek permission from any personnel appearing in images.

IV. Policies for re-use, redistribution

Our default policy is that data will be available for re-use and re-distribution under the appropriate Creative Commons License (http://creativecommons.org/licenses/), to enable maximum dissemination and use. Commercial re-use of project data, algorithms, and software may be subject to intellectual property policies at the respective partner institutions, at the discretion of each institution.

V. Plans for archiving & preservation

We will follow best practices for data management and preservation (e.g. DataOne Best Practices, http://www.dataone.org/best-practices). The CI Team will develop standard operating procedures for data management to ensure that data will be preserved in a way that maintains maximum value. In addition to efforts in metadata standards described above, we will develop and document data quality assurance and quality control plans, procedures for efficient data transfer from field devices, standardized descriptive file naming systems and file directory structures, and data backup policies.

Large amounts of data will be produced, and its protection will be critical, yet challenging. All data will be backed up on either servers available to investigators at their institutions (e.g. MU, Danforth, WUSTL) or on the proposed EPSCoR system, which will include a disk backup system. The CI Team will work with each investigator to develop a plan for backup and recovery. Backups will be documented either by the CI team for the central EPSCoR system, or by individual investigators, who will provide documentation to the CI Team. Timely submission of raw data to repositories will provide an additional layer of safety.

The CI Team will work with investigators to identify appropriate repositories. Sequencing data will be submitted to the NCBI Sequence Read Archive. iPlant will serve as a repository for plant image and phenotyping data. Other repositories will be identified through resources such as the Open Access Directory (http://oad.simmons.edu/oadwiki/Data_repositories), the Ohio State University science repository (http://library.osu.edu/find/subjects/science-data/) and the Registry of Research Data Repositories (http://www.re3data.org). Metadata, code, small datasets and links to large datasets used in publications will be archived as supplements or in repositories such as Dryad (http://datadryad.org/) or MOspace (http://libraryguides.missouri.edu/MOspace), the UM System institutional repository.