TERRA-REF Documentation

Introduction

About this book

This book describes the TERRA-REF data collection, computing, and analysis pipelines. The following links provide quick access to:

- Available Data
- How to access data
- Hands on tutorials (external)

About TERRA-REF

The ARPA-E-funded Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform (TERRA-REF) program aims to transform plant breeding by using remote sensing to quantify plant traits such as plant architecture, carbon uptake, tissue chemistry, water use, and other features to predict the yield potential and stress resistance of 300+ diverse Sorghum lines.

The data storage and computing system provides researchers with access to the reference phenotyping data and analytics resources using a high performance computing environment. The reference phenotyping data includes direct measurements and sensor observations, derived plant phenotypes, and genetic and genomic data.

Our objectives are to ensure that the software and data in the reference data and computing pipeline are interoperable, reusable, extensible, and understandable. Providing clear definitions of common formats will make it easier to analyze and exchange data and results.

We have a large and diverse team - and an even larger community of contributors. See terraref.org/team.

Users

This documentation is intended to enable a wide variety of user-groups to use and contribute to the datasets and tools contained in the TERRA-REF platform, including:

- **Plant Physiologists** interested in finding data and contributing ***phenotyping protocols***. Relevant sections include the scientific objectives and protocols.
- **Computer/Data Scientists** looking for **computer vision** and **machine learning** problems. See a complete list of data products and how to access data.
- **Engineers, Developers and Institutional IT** looking for **infrastructure** required to turn big data into scientific understanding. See an overview of the software used, the developer manual and technical documentation. **Note:** ongoing pipeline development has migrated to https://agpipeline.github.io/.

Communication

Core Communication Tools

- Slack (**signup**) terra-ref.slack.com
- We have a **Github** organization at github.com/terraref that includes repositories containing algorithms as well as discussions, documentation, and other tools. Key repositories include:
  - **Data products** repository **github.com/terraref/reference-data**
    - discussions, bug reports, and feature requests
  - **Computational Pipeline** repository **github.com/terraref/computing-pipeline**
    - discussions, bug reports, and feature request
- **Website**: terraref.org
- **Documentation** docs.terraref.org
- **Collaborative Manuscripts**
- **Email:**
  - David LeBauer (Data / Computing Lead): dlebauer@arizona.edu
  - Nadia Shakoor (Project Director): nshakoor@danforthcenter.org
  - Todd Mockler (Principal Investigator): tmockler@danforthcenter.org
  - The entire team: terraref.org/team

Scientific Objectives

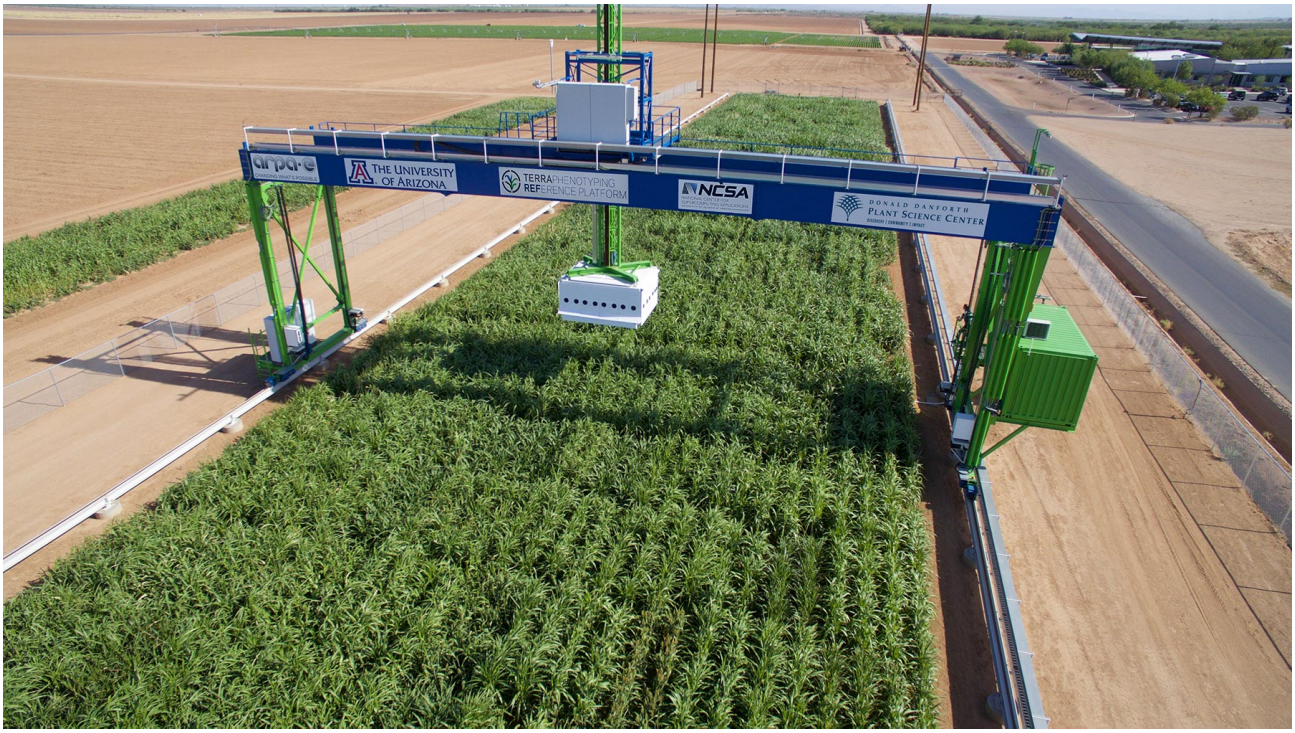
Combining advanced sensing with novel analytical approaches to accelerate breeding

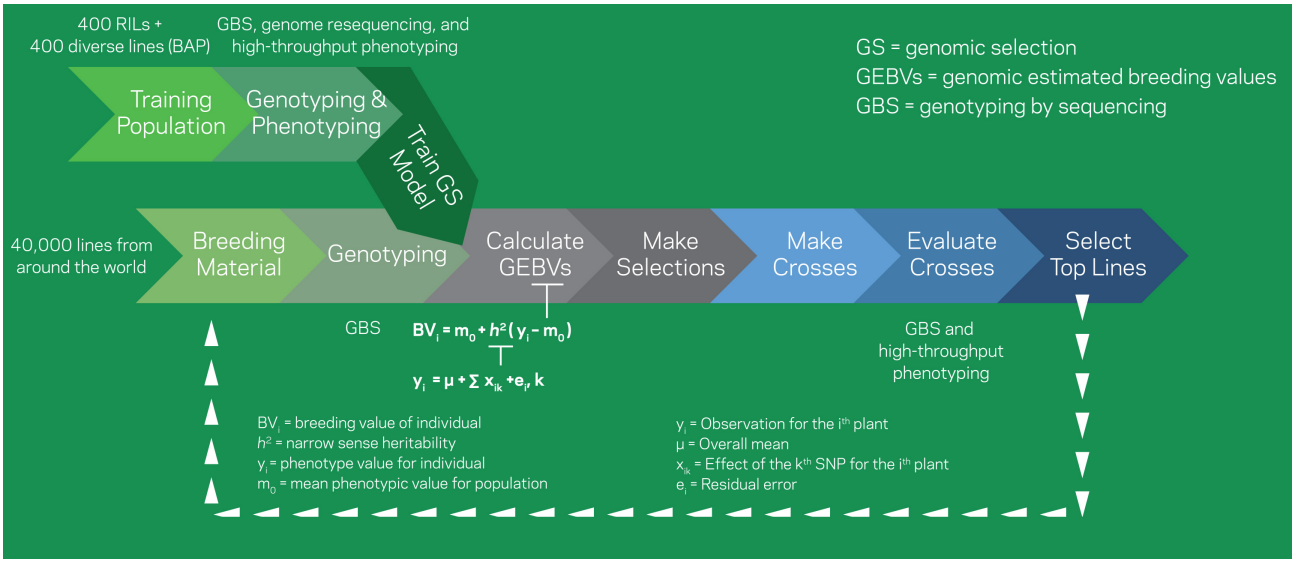
The ARPA-E-funded Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform (TERRA-REF) program aims to transform plant breeding by using remote sensing to quantify plant traits such as plant architecture, carbon uptake, tissue chemistry, water use, and other features to predict the yield potential and stress resistance of 400+ diverse Sorghum lines.

Conducting high-throughput phenotyping to connect discoveries to field performance

Breeding is currently limited by the speed at which phenotypes can be measured, and the information that can be extracted from these measurements. Current instruments used to quantify plant traits do not scale to the thousands or tens of thousands of individual plants that need to be evaluated in a breeding program. The TERRA-REF field scanner system scans over 1 acre of plants, collecting thousands of daily measurements throughout the growing season that are used to determine plant phenotypes and inform breeding decisions.

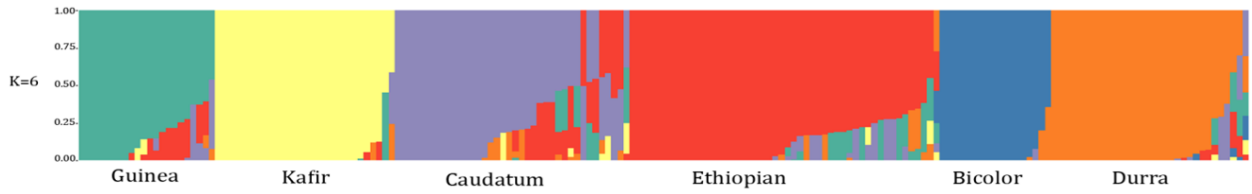
The field level phenotypic data combined with the genomic data is helping us to identify the differences between each line and the reference genome sequence for sorghum. We are using bioinformatics and quantitative genetics to characterize the observed genetic variation and identify genomic regions controlling biomass, plant architecture, and photosynthetic traits.





Using large-scale genome sequencing to drive phenotype-genotype associations and gene discovery

There is enormous potential for sorghum crop improvement. There are 50,000 sorghum accessions in the U.S. germplasm collection and most are unused and unstudied. TERRA-REF is analyzing a sorghum bioenergy association panel (BAP) that includes diverse sweet and biomass lines from all five sorghum races. The BAP captures geographic, racial, and genomic diversity.



TERRA-REF has already sequenced 384 of the lines with an average sequence coverage of 20x per line. Genome-wide association studies (GWAS) are now underway.

Providing reference quality data as a community resource

TERRA-REF is developing a data storage and computing system that provides researchers with access to all of the 'raw' data and derived plant phenotypes (traits). Data from sensors at a variety of locations across the US will be transferred to one location.

The reference data will facilitate data sharing and re-use of data by providing metadata, provenance for derived data sets, and standardized data processing workflows. It will include geospatial infrastructure for efficiently querying and transforming key datasets and tools that enable researchers to access, archive, use, and contribute data products. The technical documentation for this data pipeline is detailed in this book.

The data storage and computing system provides researchers with access to the reference phenotyping data and analytics resources using a high performance computing environment. The reference phenotyping data includes direct measurements and sensor observations, derived plant phenotypes, and genetic and genomic data.

Our objectives are to ensure that the software and data in the reference data and computing pipeline are interoperable, reusable, extensible, and understandable. Providing clear definitions of common formats will make it easier to analyze and exchange data and results.

Experimental Design

Overview

Phenotyping

The TERRA-REF project is collecting phenotype data from 852 sorghum genotypes grown in three locations:

- Maricopa Agricultural Center (MAC), Maricopa, Arizona
- Kansas State University (KSU), Ashland, KS
- Donald Danforth Plant Science Center, Missouri

Data source	MAC	KSU	Danforth
Scanalyzer Field System	X		
Scanalyzer 3D System			X
Phenotractor	X	X	
UAV (drone)	X	X	
Manual field data	X	X	

Genotyping

Whole genome resequencing is being carried out on ~400 sorghum accessions to understand the landscape of genetic variation in the selected germplasm and enable high-resolution mapping of bioenergy traits with genome wide association studies (GWAS). Additionally, ~200 sorghum recombinant inbred lines (RILs) will be characterized with ~400,000 genetic markers using genotyping-by-sequencing (Morris et al., 2013) for trait dissection in the RIL population and testcross hybrids of the RIL population.

The Maricopa Agricultural Center (MAC)

The Maricopa field site is located at the the University of Arizona Maricopa Agricultural Center and USDA Arid Land Research Station in Maricopa, Arizona.

Season 1 sorghum (April - July 2016)

Three hundred thirty one lines were planted in Season 1.

Planting maps

- Under the scanner system

Planting Design

Under scanner system

Experiment	Reps	Treatments	Experimental design
BAP	3	30 lines (12 PS, 12 sweet, 6 grain)	RCB with sorghum types nested in groups
Night illumination	3	5 illumination levels x 2 PS lines (with check line separating illumination levels)	RCB
Row #	3	6 adjacent plot scenarios: 3 lines (forage, sweet, PS) x 2 sides (east or west)	RCB but not balanced with all treatments in all reps
Biomass	3	5 sampling times x 3 lines (forage, sweet, PS)	RCB with sampling time as a repeated measure
Density	3	3 densities (5, 15, 30 cm) x 3 lines (forage, sweet, PS)	RCB

RILs	3	130 RILs plus 10 repeats of a single line/rep	Incomplete Block (row-column alpha lattice design)
Uniformity	17	2 lines (forage, PS)	None - Same line planted in single range

Season 2 sorghum (August - November 2016, unpublished)

One hundred and seventy-six lines were planted in Season 2.

Planting map

Under the scanner system

Planting Design

Under scanner system - same as season 1

Automated Phenotpying

same as season 1

Manually Collected Field Data

plant heights managements emergence vigor emergence final stand counts node and tiller counts on marked plants leaf length and width on marked plants, one date

Season 3 Durum wheat

Experimental design

Automatically and manually collected field data plan

Season 4

Season 5

Season 6

Season 7

Season 8

Season 9

Controlled Environment Phenotyping

Donald Danforth Plant Science Center, Missouri

The Automated controlled-environment phenotyping at the Donald Danforth Plant Science Center Bellwether Foundation Phenotyping Facility

The Bellwether Foundation Phenotyping Facility is a climate controlled 70 m2 growth house with a conveyor belt system for moving plants to and from fluorescence, color, and near infrared imaging cabinets. This automated, high-throughput platform allows repeated non-destructive time-series image capture and multi-parametric analysis of 1,140 plants in a single experiment. You can read more about the Danforth Plant Sciences Center Bellwether Foundation Phenotyping Facility on the DDPSC website.

The Scanalyzer 3D platform at the Bellwether Foundation Phenotyping Facility at the Donald Danforth Plant Science Center consists of multiple digital imaging chambers connected to the Conviron growth house by a conveyor belt system, resulting in a continuous imaging loop. Plants are imaged from the top and/or multiple sides, followed by digital construction of images for analysis.

- RGB imaging allows visualization and quantification of plant color and structural morphology, such as leaf area, stem diameter and plant height.
- NIR imaging enables visualization of water distribution in plants in the near infrared spectrum of 900–1700 nm.
- Fluorescent imaging uses red light excitation to visualize chlorophyll fluorescence between 680 – 900 nm. The system is equipped with a dark adaptation tunnel preceding the fluorescent imaging chamber, allowing the analysis of photosystem II efficiency.

The LemnaTec software suite is used to program and control the Scanalyzer platform, analyze the digital images and mine resulting data. Data and images are saved and stored on a secure server for further review or reanalysis.

Experiments LT1A (TM015) and LT1B (TM016)

Duration: 10 days on LemnaTec platform

Experimental Design:

- 3 replicates of 190 BAP lines were grown in a randomized complete block design
- Watering regimes = 30% FC and 100% FC
- Drought conditions were imposed 10 days after planting
- Plants were imaged daily for 10 days (11-20 DAP) and sampled at 20 days after planting
- Experiment was repeated twice to phenotype the full BAP (Reps 1A and 1B)

Genomics

Whole genome resequencing was carried out on ~400 sorghum accessions to understand the landscape of genetic variation in the selected germplasm and enable high-resolution mapping of bioenergy traits with genome wide association studies (GWAS). Additionally, ~200 sorghum recombinant inbred lines (RILs) were characterized with ~400,000 genetic markers using genotyping-by-sequencing (Morris et al., 2013) for trait dissection in the RIL population and testcross hybrids of the RIL population.

Whole-genome resequencing

Experimental Design:

- 384 BAP samples were sequenced to an average depth of ~25x.
  - List of BAP accessions in BETYdb
- Shotgun sequencing (127-bp paired-end) was done using an Illumina X10 instrument at the HudsonAlpha Institute for Biotechnology.
- Variant calling was done using a computational pipeline at the Danforth Center.
- Data were aligned against the BTx623 reference genotype
  - NCBI https://www.ncbi.nlm.nih.gov/assembly/GCF_000003195.3/
  - Phytozome https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sbicolor

---

Genotyping-by-sequencing

Experimental Design:

- 768 RIL samples were sequenced using a GBS approach.

See Also

- Genomics Protocols
- Genomics Data Availability

Protocols

The following protocols have been contributed by TERRA-REF team members:

- **Field Scanner** - Coming 2017
- **Genomics** - Coming 2017
- **Manually Collected Field Data**
- **Phenotractor**
- **UAV**

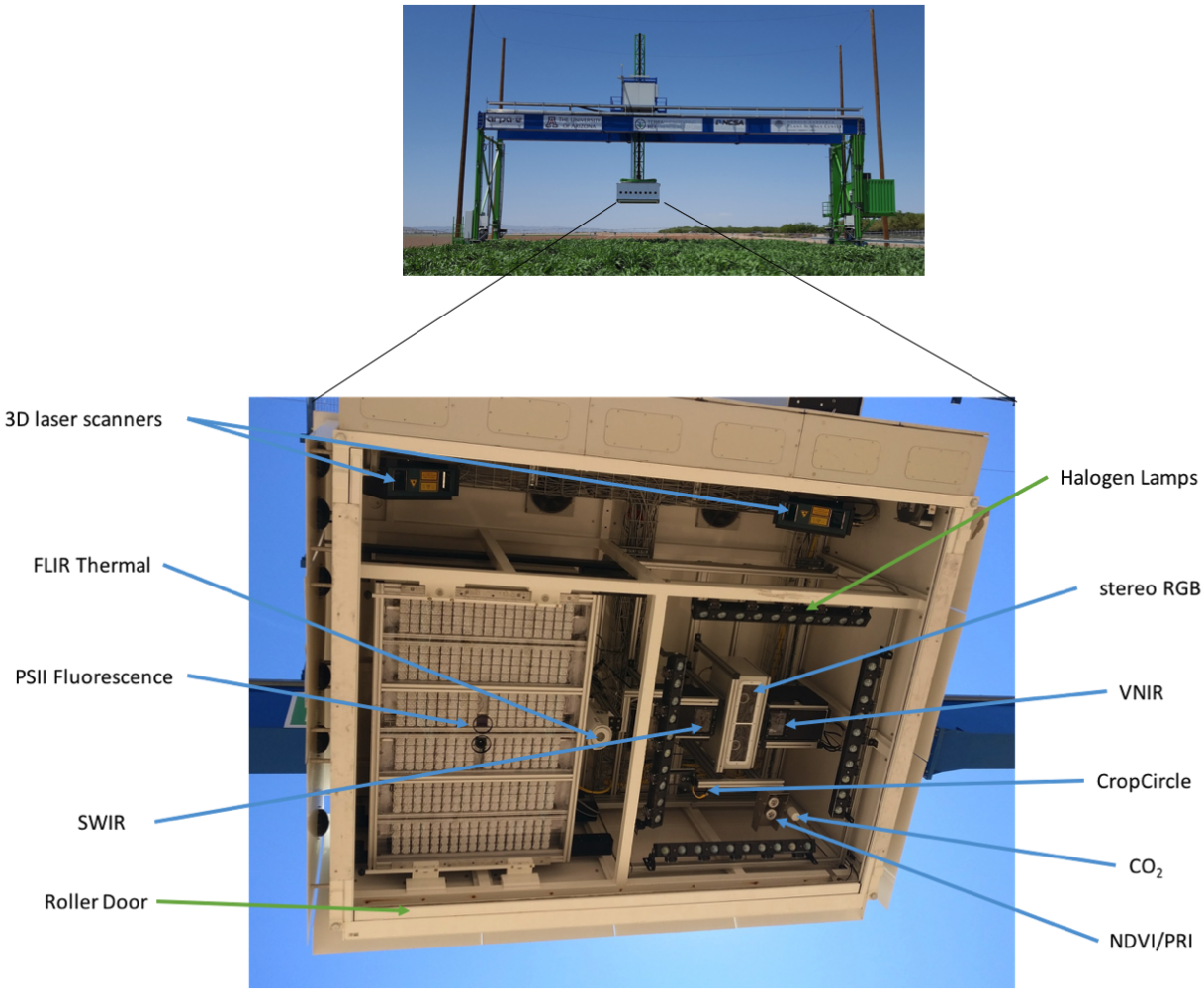
Field Scanner

This section describes the sensors deployed on the Lemnatec Field Scanner in Maricopa, AZ. Device and sensor information, including technical specifications, calibration data, and calibration targets are stored in the TERRA REF Clowder database.

The Maricopa field site is located at the the University of Arizona Maricopa Agricultural Center and USDA Arid Land Research Station in Maricopa, Arizona. At this site, we have deployed the following phenotyping platforms.

The **Lemnatec Scanalyzer Field System** is the largest field crop analytics robot in the world. This high-throughput phenotyping field-scanning robot has a 30-ton steel gantry that autonomously moves along two 200-meter steel rails while continuously imaging the crops growing below it with a diverse array of cameras and sensors.

Twelve sensors are attached to the gantry system. Detailed information for each sensor including name, variable measured, and field of view are described in the table below, with links to more detailed specifications.



Sensor Name	Model and Spec Sheet	Field of View	Pixel dimension (mm) at 2m	Technical Specifications	Notes
Imaging Sensors					
Stereo RGB Camera	GT3300 (web) (specs)	53°	0.305 x 0.315		
Laser Scanner	Custom Frauhofer 3D (specs)	0.5m width	1.0 x 0.4	Laser Power: 2000 mW Laser class: 3B	
VNIR Hyperspectral Imager	Inspector VNIR (specs)	21°	0.6	380-1000 nm @ 2/3 nm resolution	

SWIR Hyperspectral Imager	Inspector SWIR (specs)	44.5°	1.4 x 1.0	900-2500 nm @ 12 nm resolution	
Thermal Infrared	FLIR SC 615 (web) (specs)	25° x 19°	2.3	Thermal sensitivity: <50mK @ +30°C Range: -40°C to +150°C	
PSII Fluorescence Response	Lemnatec PS II (specs)	25° x 19°	1.38 x 1.35	Illumination 4000umol/m2/s Wavelength: 635nm	
Multi-spectral Radiometers					
Dedicated NDVI	Skye Multispectral Radiometer (web) (specs)			CH1: 650nm; CH2 800. Bandwidth: 10 nm	1 down, 1 up
Dedicated PRI Sensor	Skye Radiometer (web) (specs)			CH1: 531nm +/- 3nm	PRI = Photochemi Reflectance Index
PAR Sensor	Quantum SQ-300	180°		Spectral Range 410 to 655 nm	
VNIR Spectroradioometer	Ocean Optics STS-Vis			Range: 337-824 nm @ 1/2 nm	
VNIR+SWIR Spectrometer	Spectral Evolution PSR+			Range 350-2500nm @3-8 nm	Installed 20

Active Reflectance	Crop Circle ACS430P	Bands: 670, 730, 780 nm	
Environmental Sensors			
Environmental Sensors	Thies ClimaSensorD (web) (specs)	Wind Speed Wind Direction Air Temperature Relative Humidity Air Pressure Light Precipitation	Wind: 0 - 60m/s Wind directi 0 – 360° Air temperature -30°C – 70°(Relative Humidity: 0 100% Air pressure: 300 – 1100hPa Lightness: 0 150kLux
Open Path CO2 Sensor	GMP 343	CO2 Concentration Range: 0-1000 ppm	

Sensor Calibration

This section describes sensor calibration processes and how to access additional information about specific calibration protocols, calibration targets, and associated reference data.

LemnaTec Field Scanalyzer

Calibration protocols

Calibration protocols have been defined by LemnaTec in cooperation with vendors and the TERRA-REF Sensor Steering Committee. Draft calibration protocols are currently in Google Drive and have been incorporated into the LemnaTec Scanalyzer Field sensor documentation.

A detailed calibration process is also provided for the Hyperspectral sensors, with further information below.

Calibration targets

The following calibration targets are available:

- LabSphere Spectralon Diffuse Color Targets
- SphereOptics Zenith Polymer diffuse reflectance standards
- Aluminum 3D test object

Sensor Calibration

Environmental sensor calibration

The environmental sensor has been calibrated by LemnaTec. The output of the spectrometer is raw counts, users will need to use the calibration files to convert to units of

μW m-2 s-1, taking into account the bandwidth of the chip (0.4nm) if converting to μmol m-2 s-1.

Calibration reference data is available via Globus `/sites/ua-mac/EnvironmentLogger/CalibrationData` or in Github Calibrations.zip

Hyperspectral calibration

Sources:

- Convert hyperspectral exposure image to reflectance
- Hyperspectral calibration protocols

For the SWIR and VNIR sensors, factory calibration is repeated each year using the calibration lamp provided by Headwall. To convert the hyperspectral exposure image to reflectance requires the wavelength-dependent, factory calibrated reflectance of the spectralon at all VNIR and SWIR wavelengths and a good image of a spectralon panel from each camera. This includes periodic measurements of a white spectralon reflectance panel run with 20ms exposure to match panel calibration.

Dark reference measurement:

- VNIR
  - Dark measurement for VNIR camera is taken at exposure times 20, 25, 30, 35, 40, 45, 50, 55ms.
  - Data is in the same hypercube format with 180-200 lines, 955 bands, and 1600 pixel samples.
  - Data is available on Globus in /gantry_data/VNIR-DarkRef/ or via Google Drive.
  - Measurement was done using Headwall software, so there is no LemnaTec json file.
  - The name of the folder is the exposure time. "current setting exposure" is showing the exposure time in ms.
  - Custom workflow to process the calibration files.
- SWIR;
  - Dark counts handled internally, so no calibration files are necessary.

White reference measurement:

- VNIR
  - White measurement for VNIR camera is taken at exposure times 20, 25, 30, 25, 40, 25, 50, 55 ms.
  - The name of the folder is the exposure time. Data are 1600 sample, 955 bands and 268-298 lines. White reference is located in the lines between 60 to 100 and in the samples between 600 to 1000.
  - Data is available via Google Drive.
    The white reference scans was done at around 1pm (one hour after solar noon). I don't see the saturation with 20ms and 25ms exposure time.
  - For the calibration, this needs to be subtracted from the dark current in the same sample, band and exposure time.
  - In the following file, I stored an extra file named "CorrectedWhite_raw". This file includes only a single white pixel(one line, one sample) in 955 bands for each exposure time. Data is stored in the similar format but it doesnot include any extra files like frameIndex, image, header ,..
    https:\/\/drive.google.com\/file\/d\/0ByXIAClmwxA7dVNHa3pTYkFjdWc\/view?usp=sharing
    Let me know if you have issue with opening the files.

Stereo 3D height scanner

LemnaTec applied calibration matrix to the 3D scanners.

UAV calibration

Source: https://github.com/terraref/computing-pipeline/issues/185

- There are calibrated reference panels and blackbody images taken with UAV sensors before and\/or after the each flight mission.
- There are also 4 white,grey and black panels laid on the ground during the flight. Knowing the proprieties of these targets would helps us radiometrically correct the UAV images.
- What are the reflectance properties of calibrated reference panels for multispectral camera?
- What are the thermal properties of reference target for thermal camera?

- What are the reflectance properties of the reference panels laid on the ground during the flight?
- Is there any other ground truth data collected during the flight for aerial data processing, such as surface reflectance, temperature and other environmental data? These type of data would be helpful for further atmospheric correction.
- There are two sets of reference reflectance panels: one that PDS uses, it is small, PDS will need to provide the specs; the second set consists of 4 8m x 8m canvas tarps, nominally 4%, 8%, 48% and 64% reflectance across vnir bands.
- We have data from an ASD spectrometer on many but not all flight days that can be used to give the most accurate actual reflectances for each. Kelly Thorp can provide the numbers. The tarps are old and the dark targets are more reflective than nominal and light targets darker than nominal.
- The thermal target is a passive black body, I dont know the surface emissivity, it is around 0.97. There are thermistors in the back of the metal plate to provide physical temperature of the body. The black body is stored in a wood box, insulated, to dampen thermal variations. Id guess it is accurate to 2C.
- There is a met station on farm for air temperature, humidity, wind speed, wind direction, solar radiation. we have a sun photometer that can be used for atmospheric water vapor content but currently dont deploy it routinely.

Halogen spectrum

No per-wavelength analysis of light produced by the halogen lights is available from the vendor for Showtec 240V\/75W. Measurements are available for a similar halogen bulb Philips Twistline Halogen 230V 50W 18072 in Github: MeasurementPhilipsHalogenSpot.xlsx.

Spectral response data

Relative spectral response data is available for the following sensors:

- NDVI
- PRI
- PAR

Calibration data

Where available, per device calibration certificates are included in the Device and Sensor information collections.

Hyperspectral Data

The TERRA hyperspectral data pipeline processes imagery from hyperspectral camera, and ancillary metadata. The pipeline converts the "raw" ENVI-format imagery into netCDF4/HDF5 format with (currently) lossless compression that reduces their size by ~20%. The pipeline also adds suitable ancillary metadata to make the netCDF image files truly self-describing. At the end of the pipeline, the files are typically [ready for xxx]/[uploaded to yyy]/[zzz].

Installation

Software dependencies

The pipeline currently depends on three pre-requisites: *_[_netCDF Operators (NCO)*] (http://nco.sf.net)_._ Python netCDF4.

Pipeline source code

Once the pre-requisite libraries above have been installed, the pipeline itself may be installed by checking-out the TERRAREF computing-pipeline repository. The relevant scripts for hyperspectral imagery are:

- Main script terraref.sh* JSON metadata->netCDF4 script JsonDealer.py

Setup

The pipeline works with input from any location (directories, files, or stdin). Supply the raw image filename(s) (e.g., meat_raw), and the pipeline derives the ancillary filename(s) from this (e.g., meat_raw.hdr, meat_metadata.json). When specifying a directory without a specifice filename, the pipeline processes all files with the suffix "_raw".

```
shmkdir ~/terrarefcd ~/terrarefgit clone git@github.com:terraref/computing-pipeline.gitgit clone git@github.com:terraref/documentation.git
```

Run the Hyperspectral Pipeline

```
shterraref.sh -i ${DATA}/terraref/foo_raw -O ${DATA}/terrarefterraref.sh -I
/projects/arpae/terraref/raw_data/lemnatec_field -O
/projects/arpae/terraref/outputs/lemnatec_field
```

Controlled Environment Protocols

Authors: Mockler Lab

Abstract

Automated VIS and NIR imaging in a controlled growth environment

Materials

- ProMix BRK20 + 14-14-14 Osmocote pots
- pre-filled by Hummert Sorghum seed

Equipment

- Conviron Growth House
- LemnaTec moving field conveyor belt system
- Scanalyzer 3D platform

Procedures

Planting

- Plant directly into phenotyping pots □

Chamber Conditions

Pre-growth (11 days) and Phenotying (11 days)

- 14 hour photoperiod
- 32oC day/22oC night temperature
- 60% relative humidity
- 700 umol/m2/s light

Watering Conditions

- Prior to phenotyping, plants watered daily
- The first night after loading, plants watered 1× by treatment group to 100% field capacity (fc)
- Days 2 – 12, plants watered 2× daily by treatment group (100% or 30% FC) to target weight

Automation

- Left shift lane rotation within each GH, during overnight watering jobs
- VIS (TV and 2 x SV), NIR (TV and 2 x SV) imaging daily

Recipes

- **Field capacity** = 200% GWC (200 g water/100 g soil), based upon extensive GWC testing done by Skyler Mitchell
- **Target weight** (fc) = [(water weight at % fc) + [(average weight of carrier/saucer) + (dry soil weight) + (pot weight)]
- **Water weight** at 100% fc = dry soil weight * (%GWC/100)
- **Water weight** at 30% fc = water weight at 100% fc * 0.30

References

Manual Field Data Protocols

Abstract

Materials

- barcode scanning protractor
- barcode scanning ruler
- ceptometer (Decagon AccuPAR LP-80)
- digital caliper
- drying oven
- forage chopper
- hand shears
- infrared thermometer
- juice extractor
- leaf area meter (Li-Cor 3100, Li-Cor Inc.)
- leaf porometer (SC-1 Leaf Porometer, Decagon Devices)
- leaf punch
- meter stick
- paper bags
- portable photosynthesis system (Li-Cor 6400, Li-Cor Inc.)
- scale
- SPAD Meter (SPAD 502 Plus Chlorophyll Meter, Minolta)
- spray paint

Equipment

Procedures

Variable

Canopy Height	Canopy height for single row of central 2 data rows of 4-row plot. Measured in cm using meter stick, taken at the height representing the plot 'potential', ignoring stunted plants. The canopy height was measured as the height of the foliage (not the inflorescence) at the general top of the canopy where the upper leaves bend and/or establish a canopy surface that would support a very light horizontal object (imagining a light sheet of rigid plastic foam), discounting rare or exceptional leaves in the upper-most 2 or 3 percentile.
Panicle Height	Height of the top of the inflorescence panicle for single central data row of 4-row plot, when panicle extends notably above canopy height.
Seedling Vigor and Emergence	Count the number of emerging seedlings at about 20% emergence, and then repeat every other day until final stand is achieved. A seedling is defined as emerged when the coleoptile is visible above the soil surface. Final stand is defined as when a similar count +/- 5% is achieved on successive counts 1-2 days apart. Count seedlings in the entire plot. Two Alternatives 1. Explicitly count number of plants emerged 2. For each plot, assess % germination in categories (e.g. [0,20], [20,40], …) This is the standard method
Canopy closure and leaf area index	Sunfleck ceptometer readings will be taken at least monthly to determine radiation interception and canopy closure. Using e.g. Decagon AccuPAR LP-80. Leaf area index will be calculated using Beer's Law for light extinction. A total of 5 readings will be taken per plot and averaged. Readings will be taken on clear days. Incident light will be measured at least once per rep. NDVI will also be measured weekly using a tractor mounted unit until the tractor can no longer navigate through the field due to the height of the crop. References:Prometheus Wiki http://prometheuswiki.publish.csiro.au/tiki-index.php?page=Canopy+light+interception+assessment+-+from+DC20
Leaf Architecture / Leaf erectness	Barcode scanning protractor is used to measure youngest fully emerged leaf

Leaf Width	Barcode scanning ruler measured at the widest part of the leaf
Stem number	Manually count the total number of stems in the plot will be counted bi-weekly after thinning for all plants in the plot.
Stem diameter	Stem diameter for each of 10 plants per plot will be measured with a digital caliper at 10 and 150 cm every month. For each plant take a few diameter samples and record the most common value. Use a black sharpie to mark the location at which the sample was taken.
Canopy Height	An "eyeball" estimate of plant height for the entire plot will be taken weekly beginning at the 5-leaf stage. Canopy height, view the canopy horizontally with a measuring stick, taking the height where a light piece of foam would rest on the canopy. Estimate the median height of healthy standing plots, ignoring plants that look really bad (e.g. are lodged). For method development: on subset of plots (10), capture the distribution of heights, e.g. max, min, median, upper and lower quantiles.

Lodging	There are three measures: 1. Percent lodging 0-100 scale 2. Lodging severity 0-100 scale 3. Lodging score 0-100 scale 4. Whether this is stalk or root lodging (categorical 'root', 'stalk') A lodging score will be taken weekly once lodging is observed. The lodging score will be recorded as a percentage and is a combination of the fraction of the plants lodged and the severity of lodging. For example, if 50% of the plants are 50% lodged, then the lodging score would be 25%. The severity of lodging is determined by how far the plants are leaning from vertical. If a plant is laying on the ground the severity of lodging is 100%. If a plant is leaning 45 degrees from vertical, then the severity of lodging is 50%. How to differentiate between stalk lodging and root lodging: scoring 'lodging' implies diagnosing a cause of inclined stems. A better approach may be a visual estimate of a range, with an optional note for root or shoot lodging. Done as deflection from vertical, this might look like:Min_angle Max_angle Loding_type0 1010 4530 60 R20 40 S…Where R = root lodging, S = stem lodging. Since stems are usually curved, the question remains of what reference height to consider?
Above-ground yield	Alleyways will be trimmed by hand with a weed whacker with a blade to accommodate space required between plots for a 2-row forage chopper. Actual plot length will be measured from the first to last stalk cut by the forage chopper. The stalks trimmed by hand will be spray painted to delineate them from stalks in the harvest area. The chopped forage will be weighed in a bag and a 2-quart sample removed for moisture and quality analysis. The sample will be dried in an oven at 65 C until constant weight is achieved. The dried forage will be ground and submitted for quality analysis. Sorghum Checkoff provided 1.5 pg protocol

Total biomass and tissue partitioning	Plants will be (destructively?) sampled (from west of gantry plots?) five times during the season from the 5 leaf stage through final harvest. The area sampled will be 1 meter of row. The plants will be cut off at ground level and immediately placed in a cooled ice chest for transport from the field to the laboratory where they were stored at 5°C until processing.
Allometry	Plant height will be measured from the base of the plant to the point where the top leaf blade is perpendicular to the stem. The number of stems and their average phenological stage will be recorded. Leaves will be removed from the stem at the collar and separated into green and brown leaves.
Leaf Area Index (LAI)	Leaf area of green leaves will be measured with a leaf area meter (Li-Cor 3100, Li-Cor Inc., Lincoln, NE, USA). Heads will be separated from the stems. Stem area will be estimated from stem length (without the head) x diameter. The stems, brown and green leaves, and heads will be dried separately in an oven at 65°C for 2–4 d and weighed. Leaf area index and stem area index will be calculated.
Specific Leaf Area (SLA)	Specific leaf area will be calculated by dividing green leaf area by green leaf weight.
Phenology	Phenology will be determined according to Vanderlip (1993). Before heading, developmental stages were based on the appearance of the leaf collars. After heading, phenological stages were determined based on the development of the grain. Numbers ranging from 1 (50% of plants heading) to 7 (50% of plants at physiological maturity) were assigned to designate growth stage after the vegetative period. Before heading, growth stages represent mean leaf number of all plants and not the most advanced 50% as was done after headingReference: https://www.bookstore.ksre.ksu.edu/pubs/S3.pdf
Days to flag leaf emergence	

Days to spike emergence	
Days to anthesis/flowering	Once anthesis begins, anthesis will be noted 3 times per week until anthesis ends. Anthesis is defined as when 50% of the plants have one or more anthers showing.
Maturity pattern	Once maturity begins, maturity will be noted 3 times per week until maturity ends. Maturity is defined as when 50% of the plants have reached black layer.
Moisture content	Forage moisture content will be determined at final harvest and from the biomass samples by weighing the forage before and after drying in an oven at 65 C for a minimum of 48 h. How large is the sample? ~ 1 pound in a lunchbag, 2 samples per plotHow will it be packaged / labeled?Subsamples?
Lignin content	Determined by NIRS from the moisture sample at final harvest.
BTU/DW	Determined by NIRS from the moisture sample at final harvest.
Juice extraction	Juice will be extracted from stalks from the biomass samples at final harvest using a sweet sorghum mill. The juice will be weighed and brix measured. Brix concentration in the juice – Brix will be measured in the juice extracted as described above.
Plant temperature	A hand-held infrared thermometer will be used to measure plant temperature bi-weekly. A total of 5 readings will be recorded per plot within 2 hours of solar noon.
Plant color	A Minolta SPAD meter will be used to record plant color on plants using the most recently fully expanded leaf on a bi-weekly basis.
Photosynthesis	Using LiCOR 6400, measure A-Ci and A-Q curves to estimate parameters of Collatz model of C4 photosynthesis coupled to the Ball Berry model of stomatal conductance. One reading from the youngest fully expanded leaf. These readings will be taken monthly within 2 hours of solar noon.

Transpiration/stomatal conductance	Stomatal conductance was assessed using a leaf porometer (Decagon Devices, Pullman, WA) by taking 5 readings per plot on most recently fully expanded leaves. Readings will be taken on the 12 photoperiod sensitive lines in the biomass association panel. These readings will be taken bi-weekly and within 2 hours of solar noon at least two times during the season.

References

- Pérez-Harguindeguy N., Díaz S., Garnier E., Lavorel S., Poorter H., Jaureguiberry P., Bret-Harte M. S., CornwellW. K., Craine J. M., Gurvich D. E., Urcelay C., Veneklaas E. J., Reich P. B., Poorter L., Wright I. J., Ray P., Enrico L.,Pausas J. G., de Vos A. C., Buchmann N., Funes G., Quétier F., Hodgson J. G., Thompson K., Morgan H. D., ter Steege H., van der Heijden M. G. A., Sack L., Blonder B., Poschlod P., Vaieretti M. V., Conti G., Staver A. C.,Aquino S., Cornelissen J. H. C. (2013) New handbook for standardised measurement of plant functional traits worldwide. *Australian Journal of Botany* **61** , 167–234. https://doi.org/10.1071/BT12225
- Vanderlip RL. 1993. How a sorghum plant develops. Manhattan, KS, USA: Kansas State University Cooperative Extension. Field Experiments in Crop Physiology. 2013, Jan 13. In *PrometheusWiki*. Retrieved 15:03,June 21, 2016, from http://www.publish.csiro.au/prometheuswiki/tiki-pagehistory.php?page=Field Experiments in Crop Physiology&preview=41

Photosynthesis / leaf chemistry from hyperspectral data references:

- Shawn Serbin et al - Leaf optical properties reflect variation in photosynthetic metabolism and its sensitivity to temperature 2011 J Exp Bot
- Mapping biochemistry and photosynthetic metabolism in ecosystems using imaging spectroscopy (Presentation) - Remotely estimating photosynthetic capacity, and its response to temperature, in vegetation canopies using imaging spectroscopy 2015 Remote Sensing of the Environment
- Spectroscopic determination of leaf morphological and biochemical traits for northern temperate and boreal tree species 2014 Ecological Applications

- Additional Draft Protocols are available at https://docs.google.com/document/d/1iP8b97kmOyPmETQI_aWbgV_1V6QiKYLblq1jIqXLJ84/edit#

Phenotractor Protocols

Authors: Matthew Maimaitiyiming, Wasit Wulamu, and David LeBauer

Center for Sustainability, Saint Louis University, St. Louis, MO 63108

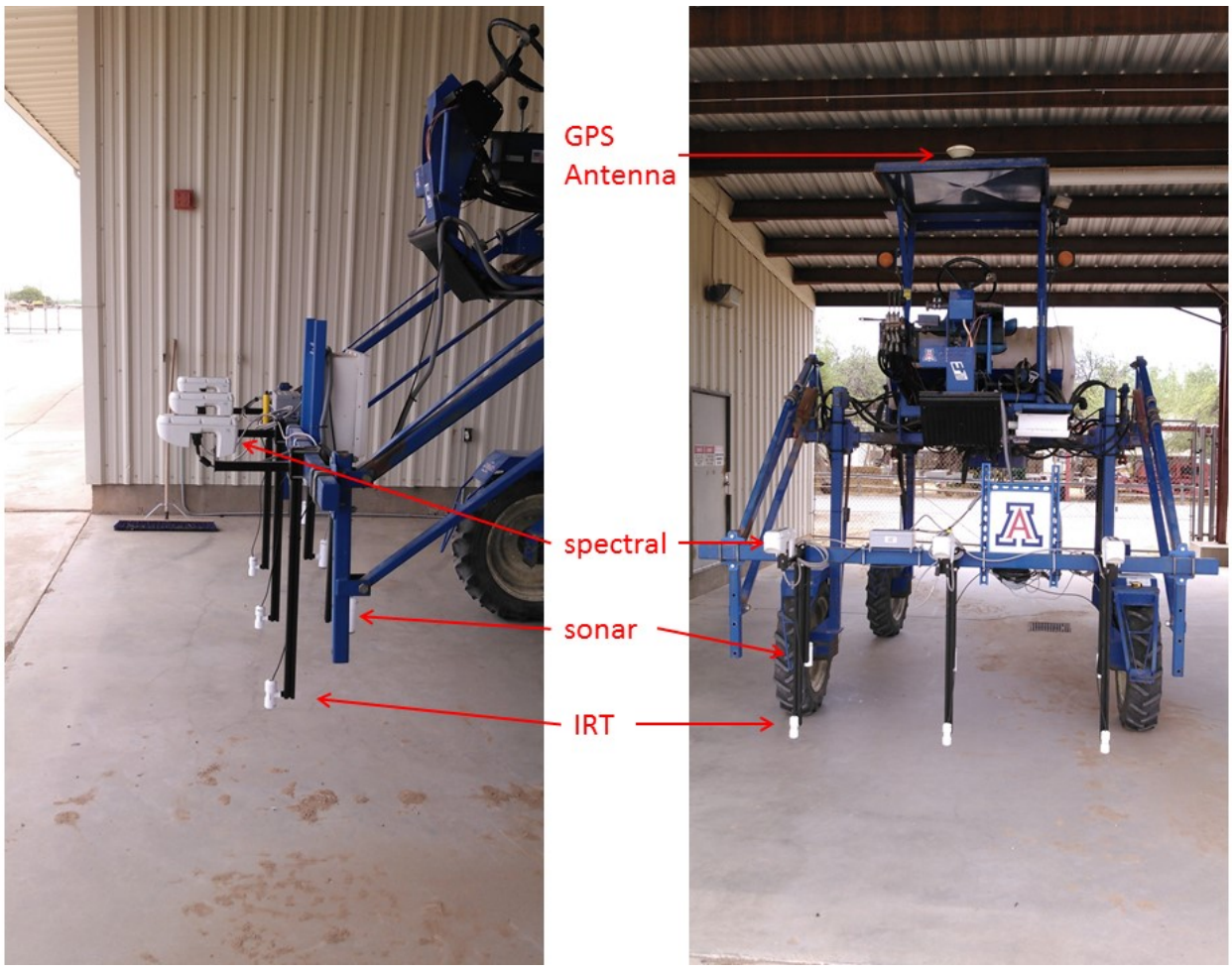
Abstract

This document provides a brief summary of methods, procedures, and workflows to process the tractor data.

Content modified from Andrade-Sanchez et al 2014.

Materials

- Tractor
- Sensors
  - Sonar Transducer
  - GreenSeeker Multispectral Radiometer
  - Infrared Thermal Sensor



Picture of Phenotractor Sensors

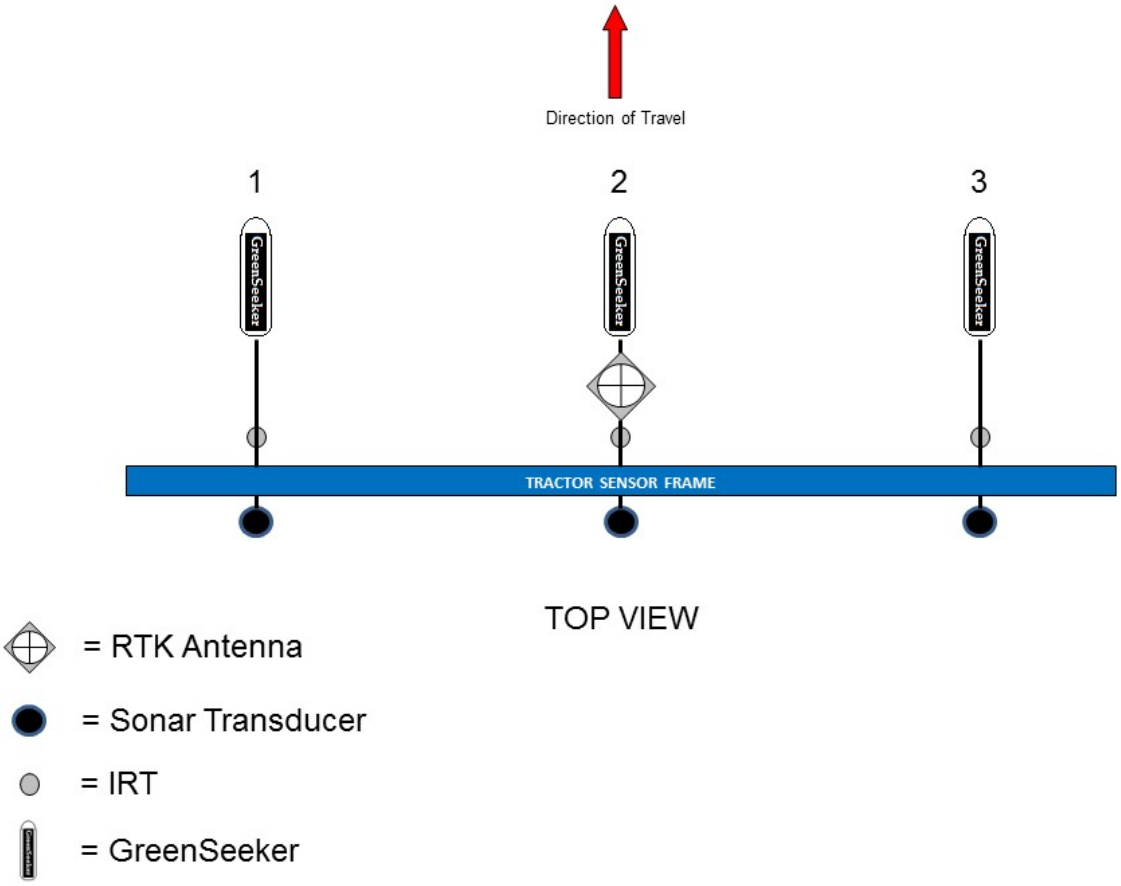


Diagram of sensor attachments

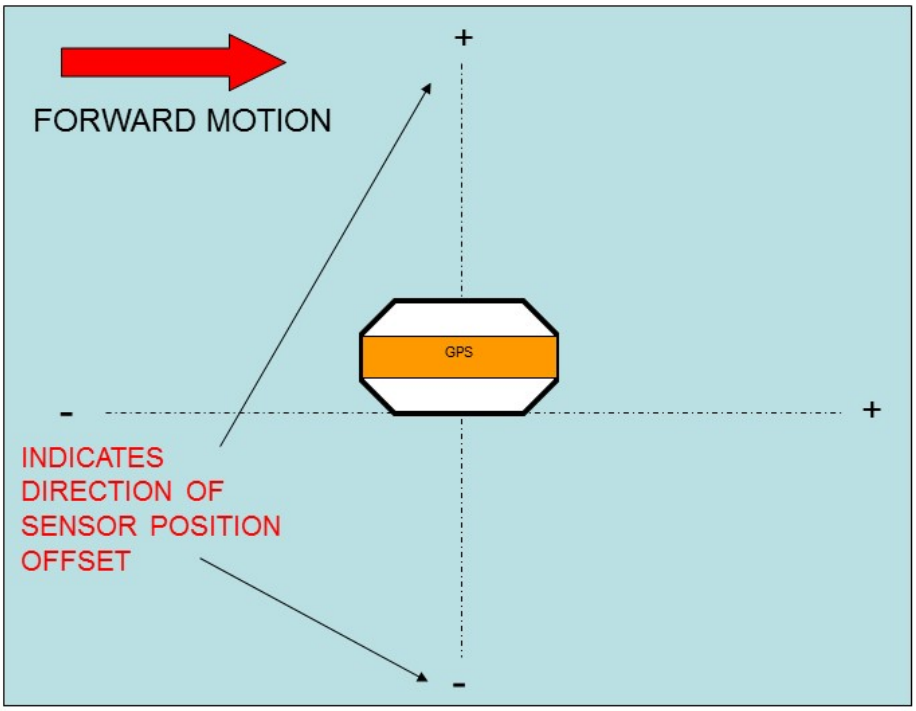


Diagram of Sensor Offset

Methods

Tractor

The Tractor-based plant phenotyping system (Phenotractor) was built on a LeeAgra 3434 DL open rider sprayer. The vehicle has a clearance of 1.93 m. A boom attached to the front end of the tractor frame holds the sensors, data loggers, and other instrumentation components including enclosure boxes and cables. The boom can be moved up and down with sensors remaining on a horizonal plane. An isolated secondary power source supplies 12-V direct current to the electronic components used for phenotyping.

Phenotractor system configuration

Sensors

The phenotractor was equipped with three types of sensors for measuring plant height, temperature and canopy spectral reflectance. A RTK GPS was installed on top of the tractor, see the figure below.

Plant Height with Sonar

The distance from canopy to sensor position was measured with a sonar proximity sensor ($S \rm{output}$, in mm). Canopy height (CH) was determined by combining sonar and GPS elevation data (expressed as meter above sea level). An elevation survey was conducted to determine a baseline reference elevation ($E\rm{ref}$) for the gantry field. CH was computed according to the following equation:

where $E_rm{s}$ is sensor elevation, which was calculated by subtracting the vertical offset between the GPS antenna and sonar sensor from GPS antenna elevation.

Thermal Sensor

An Apogee SI-121 Infrared radiometer (IRT) sensors were used measure canopy temperature and temperature values were recoded as degree Celsius (°C).

Multispectral Radiometer

Canopy spectral reflectance was measured with GreenSeeker sensors and the reflectance data were used to calculate NDVI (Normalized Difference Vegetation Index). GreenSeeker sensors record reflected light energy in near infrared (780 ± 15 nm) and red (660 ± 10 nm) portion electromagnetic spectrum from top of the canopy by using a self-illuminated light source. NDVI was calculated using following equation:

Where $\rho \rm{NIR}$ and $\rho\rm{red}$ and ρ_red represent fraction of reflected energy in near infrared and red spectral regions, respectively.

Georeferencing

Georefencing was carried out using a specially developed Quantum GIS (GGIS, www.qgis.org) plug-in by Andrade-Sanchez et al. (2014) during post processing. Latitude and longitude coordinates were converted to UTM coordinate system. Offset from the sensors to the GPS position on the tractor heading were computed and corrected. Next, the tractor data, which uses UTM Zone 12 (MAC coordinates), was transformed to EPSG:4326 (WGS84) USDA coordinates by performing a linear shifting as follows:

- Latitude: $U_y = M_y – 0.000015258894$
- Longitude: $U_x = M_x + 0.000020308287$

where U_y and U_x are latitude and longitude in USDA coordinate system, and M_y and M_x are latitude and longitude in MAC coordinate system (see section on geospatial coordinate systems). Finally, georeferenced tractor data was overlaid on the gantry field polygon and mean value for each plot/genotype was calculated using the data points that fall inside the plot polygon within ArcGIS Version 10.2 (ESRI. Redlands, CA).

References

Andrade-Sanchez, Pedro, Michael A. Gore, John T. Heun, Kelly R. Thorp, A. Elizabete Carmo-Silva, Andrew N. French, Michael E. Salvucci, and Jeffrey W. White. "Development and evaluation of a field-based high-throughput phenotyping platform." Functional Plant Biology 41, no. 1 (2014): 68-79. doi:10.1071/FP13126

UAV Protocols

Authors: Rick Ward

Abstract

Multispectral data collected during seasons 1-5 at Maricopa using small unmanned aircraft systems, i.e. UAVs. Workflow includes image capture with cameras on UAV platforms, generation of georeferenced orthomosic reflectance and index (e.g. NDVI) geotiffs, extraction of plot level statistics within qgis with the aid of polygon shape files in which plot attributes are stored. Downstream users can access the radiometric and index data from the reflectance map geotiffs directly, or from the plot level data uploads.

Materials

Platforms

- SenseFly eBee fixed-wing drone
- Hexacopter

Cameras

UAV data are collected using one of three cameras:

- 5-band MicaSense RedEdge
- 4-band + RGB Parrot Sequoia
- SenseFly thermal thermoMap

Cameras are carried singly or in tandem on the SenseFly eBee fixed-wing drone (Sequoia and thermoMap, individually only), or a hexacopter (RedEdge or Sequoia, individually or in

tandem).

Procedure

Flight

Standard flight altitude is 44m with 75% image overlap (both sequentially and laterally), and missions are programmed and managed by either Mission Planner or senseFly eMotion.

Calibration

No radiometric calibration was conducted as of Nov 5, 2016.

Analysis

Pix4D software was used to generate gray-scale orthomosaic geotiff files containing NDVI data after georegistration to the WGS84/UTM 12 N coordinate reference system using three to five 2D geo-located ground control points. These are manually matched to 5-40 images each. Ground control points for the Lemnatec Field Scanner are on the concrete pylons and were geolocated using an RTK base station maintained by the USDA-ARS at Maricopa (see section on geospatial information).

QGIS software was used to confirm geospatial alignment of NDVI geotiffs with shape files containing geolocated positions of the rail foundations. A shape file containing polygons aligning with the middle two rows of each of the 350 experimental units (for sorghum crop Aug-Nov 2016) was kindly generated by Dr. A French of USDA-ARS. Zonal Statistics in QGIS was used to calculate NDVI means for each plot polygon.

References

- MicaSense: https://www.micasense.com

- SenseFly https://www.sensefly.com
- QGIS https://www.qgis.org
- Pix4D https://www.Pix4D.com

Genomic Protocols

Genomic data includes whole-genome resequencing data from the HudsonAlpha Institute for Biotechnology, Alabama for 384 samples for accessions from the sorghum Bioenergy Association Panel (BAP) and genotyping-by-sequencing (GBS) data from Kansas State University for 768 samples from a population of sorghum recombinant inbred lines (RIL).

Danforth Center genomics pipeline

Outlined below are the steps taken to create a raw vcf file from paired end raw FASTQ files. This was done for each sequenced accession so a HTCondor DAG Workflow was written to streamline the processing of those ~200 accessions. While some cpu and memory parameters have been included within the example steps below those parameters varied from sample to sample and the workflow has been honed to accomodate that variation. This pipeline is subject to modification based on software updates and changes to software best practices.

Software versions:

- BBDuk2 version 36.67
- bwa v 0.7.12-r1039
- samtools v 1.3.1
- picard-tools-2.0.1
- GATK v3.5-0-g36282e4
- VCFtools (0.1.14)
- Tassel Version: 5.2.27

Preparing reference genome

Download Sorghum bicolor v3.1 from Phytozome

Generate:

BWA index:

```
bwa index -a bwtsw Sbicolor_313_v3.0.fa
```

fasta file index:

```
samtools faidx Sbicolor_313_v3.0.fa
```

Sequence dictionary:

```
java -jar picard.jar CreateSequenceDictionary R=Sbicolor_313_v3.0.fa O=Sbicc
```

Quality trimming and filtering of paired end reads

```
bbduk2 in=SampleA_R1.fastq in2=SampleA_R2.fastq out=SampleA_R1.PE.fastq.gz
  out2=SampleA_R2.PE.fastq.gz k=23 mink=11 hdist=1 tpe tbo qtrim=rl trimq=
  minlen=20 rref=adapters_file.fa lref=adapters_file.fa
```

Aligning reads to the reference

```
bwa mem -M \
  -R “@RG\tIDSAMPLEA_RG1\tPL:illumina\tPU:FLOWCELL_BARCODE.LANE.SAMPLE_BAR
  Sbicolor_313_v3.0.fa SampleA_R1.PE.fastq.gz SampleA_R2.PE.fastq.gz > SAM
```

Convert and Sort bam

```
Samtools view -bS SAMPLEA.bwa.sam | samtools sort - SAMPLEA.bwa.sorted
```

Mark Duplicates

```
java -Xmx8g -jar picard.jar MarkDuplicates MAX_FILE_HANDLES_FOR_READ_ENDS_
  REMOVE_DUPLICATES=true INPUT=SAMPLEA.bwa.sorted.bam OUTPUT=SAMPLEA.dedup
  METRICS_FILES=SAMPLEA.dedup.metrics
```

Index bam files

```
samtools index SAMPLEA.dedup.bam
```

Find intervals to analyze

```
java -Xmx8g -jar GenomeAnalysisTK.jar -T RealignerTargetCreator \
  -R Sbicolor_313_v3.0.fa -I SAMPLEA.dedup.bam -o SAMPLEA.realignment.inte
```

Realign

```
java -Xmx8g -jar GenomeAnalysisTK.jar -T IndelRealigner -R Sbicolor_313_v3
  -I SAMPLEA.dedup.bam -targetIntervals SAMPLEA.realignment.intervals -o S
```

Variant Calling with GATK HaplotypeCaller

```
java -Xmx8g -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R Sbicolor_313_v
  -I SAMPLEA.dedup.realigned.bam --emitRefConfidence GVCF --pcr_indel_mode
  -o SAMPLEA.output.raw.snps.indels.g.vcf
```

Above this point is the workflow for the creation of the gVCF files for this project. The following additional steps were used to create the Hapmap file

Combining gVCFs with GATK CombineGVCFs

NOTE: This project has 363 gvcfs: multiple instances of CombineGVCFs, with unique subsets of gvcf files, were run in parallel to speed up this step below are examples

```
java -Xmx8g -jar GenomeAnalysisTK.jar -T CombineGVCFs -R Sbicolor_313_v3.0
  -V SAMPLEA.output.raw.snps.indels.g.vcf --variant SAMPLEB.output.raw.snp
  -V SAMPLEC.output.raw.snps.indels.g.vcf -o SamplesABC_combined_gvcfs.vcf

java -Xmx8g -jar GenomeAnalysisTK.jar -T CombineGVCFs -R Sbicolor_313_v3.0
  --variant SAMPLED.output.raw.snps.indels.g.vcf -V SAMPLEE.output.raw.snp
  -V SAMPLEF.output.raw.snps.indels.g.vcf -o SamplesDEF_combined_gvcfs.vcf
```

Joint genotyping on gVCF files with GATK GenotypeGVCFs

```
java -Xmx8g -jar GenomeAnalysisTK.jar -T GenotypeGVCFs -R Sbicolor_313_v3.
  -V SamplesABC_combined_gvcfs.vcf -V SamplesDEF_combined_gvcfs.vcf -o all
```

Applying hard SNP filters with GATK VariantFiltration

```
java -Xmx8g -jar GenomeAnalysisTK.jar -T VariantFiltration -R Sbicolor_313
  -V all_combined_Genotyped_lines.vcf -o all_combined_Genotyped_lines_filt
  --filterExpression "QD < 2.0" --filterName "QD" --filterExpression "FS >
  --filterName "FS" --filterExpression "MQ < 40.0" --filterName "MQ" --fil
  --filterName "MQRankSum" --filterExpression "ReadPosRankSum < -8.0" --fi
```

Filter and recode VCF with VCFtools

```
vcftools --vcf all_combined_Genotyped_lines_filtered.vcf --min-alleles 2 -
  --out all_combined_Genotyped_lines_vcftools.filtered.recode.vcf --max-mi
```

Adapt VCF for use with Tassel5

```
tassel-5-standalone/run_pipeline.pl -Xms75G -Xmx265G -SortGenotypeFilePlug
  -inputFile all_combined_Genotyped_lines_vcftools.filtered.recode.vcf \
  -outFile all_combined_Genotyped_lines_vcftools.filtered.recode.sorted.vc
```

Convert VCF to Hapmap with Tassel5

```
tassel-5-standalone/run_pipeline.pl -Xms75G -Xmx290G -fork1 -vcf \
  all_combined_Genotyped_lines_vcftools.filtered.recode.sorted.vcf -export
```

CoGe genomics pipeline

CoGe has integrated the tools that make up the Danforth Center's variant calling pipeline into their easy point and click GUI, allowing users to reproduce a majority of the TERRA SNP analysis. Below, we detail how to run sequence data through CoGe's system.

- Goto https://genomevolution.org/coge/ or click create an account to get started.
- If this is your initial attempt, you will need to create a Genome.
    1. Under Tools, click Load Genome or use this link.

- Under Tools, click Load Experiment or use this link.
- **Select Data:** to use the TERRA data click Community Data or choose from CoGe's other data options.
- **Select Options:** This outlines CoGe's choices for data processing and analysis. To reproduce pipeline used to create the TERRA SNPs, you can reference the exact tools and parameters used in the Danforth analysis above and enter the appropriate values into their equivalent drop downs or fields.

For the TERRA SNP the following were used:

FASTQ Format

- Read Type: Paired-end
- Encoding: phred33

Trimming

- Trimmer: BBDuk
- BBDuk parameters: k=23, mink=11, hdist=1, check mark both tpe and tbo, qtrim=rl, trimq=20, minlen=20, set trim adapters to both ends

Alignment

- Aligner: BWA-MEM
- BWA-MEM parameters: check mark -M, fill in Read Groups ID (identifier), PL (sequence platform), LB (library prep), SM (sample name)

SNP Analysis

- Check mark Enable which expands this section
- Method: GATK HaplotypeCaller (single-sample GVCF) using the default parameters but you can choose to use Realign reads around INDELS

General Options

- Checkmark both options to add results to your notebook and receive an email when pipeline has completed.

Describe Experiment: Enter an experiment name (required), your data processing version ie 1 for first time, Source if using TERRA Data, it's TERRA (required), and Genome (required and if you start typing it will find your loaded genome but be sure to verify version and id .)

Data

Overview

This user manual is divided into the following sections:

- Data Products: A summary of the available data products and the processes used to create them
- Data Access: Instructions for how to access the data products using Clowder, Globus, BETYdb, and CoGe
- Description of the scientific objectives and experimental design
- Data use policy: Information about data use and attribution
- User Tutorials: In-depth examples of how to access and use the TERRA-REF data

---

What data is available?

- Raw output from sensors deployed on the Lemnatec field scanner
    - Additional data from greenhouse systems, UAVs, and tractors have not been released, but can be accessed through our beta user program
- Manually-collected fieldbooks and associated protocols
- Derived data, including phenomics data, from computational approaches
- Genomic pipeline data

Data Products

The following table lists available TERRA-REF data products. The table will be updated as new datasets are released. Links are provided to pages with detailed information about each data product including sensor descriptions, algorithm (extractor) information, protocols, and data access instructions.

Data product	Description
3D point cloud data	3D point cloud data (LAS) of the field constructed from the Fraunhofer 3D scanner output (PLY).
Fluorescence intensity imaging	Fluorescence intensity imaging is collected using the PSII LemnaTec camera. Raw camera output is converted to (netCDF/GeoTIFF)
Hyperspectral imaging data	Hyperspectral imaging data from the SWIR and VNIR Headwall Inspector sensors are converted to netCDF output using the hyperspectral extractor.
Infrared heat imaging data	Infrared heat imaging data is collected using FLIR sensor. Raw output is converted to GeoTIFF using the FLIR extractor.
Multispectral radiometer data	Multispectral data is collected using the PRI and NDVI Skye sensors. Raw output is converted to timeseries data using the multispectral extractor.
Stereo imaging data	Stereo imaging data is collected using the Prosilica cameras. Full-color images are reconstructed in GeoTIFF format using the de-mosaic extractor. A full-field mosaic is generated using the full-field mosaic extractor.
Spectral reflectance data	Spectral reflectance is measured using a Crop Circle active crop canopy sensor

Environmental conditions	Environment conditions are collected through the CO2 sensor and Thies Clima. Raw output is converted to netCFG using the environmental-logger extractor.
Meteorological data	postGIS/netCDF
Phenotype data	Phenotype data is derived from sensor output using the PlantCV extractor and imported into BETYdb.
Genomics data	FASTQ and VCF files available via Globus
UAV and Phenotractor	Plot level data available in BETYdb

See also

- Sensor calibration
- Fieldbooks and Protocols
- Data standards
- Geospatial information

Environmental conditions

Environment conditions data is collected using the Vaisala CO2, Thies Clima weather sensors as well as lightning, irrigation, and weather data collected at the Maricopa site.

Data formats follow the Climate and Forecast (CF) conventions for variable names and units. Environmental data are stored in the Geostreams database.

Data sources

- WeatherStation coordinates are 33.074457 N, 111.975163 W
- EnvironmentLogger is on top of the gantry system and is moveable.
- Irrigation is managed at the field level. There are four regions that can be irrigated at different rates.

Data access

Level 1 data

Level 1 meteorological data is aggregated to from 1 Hz raw data to 5 minute averages or sums.

netCDF: 5s (12 per minute) observations

On Globus or Workbench you can find these data provided in both hourly and daily files. These files contain data at the original temporal resolution of 1/s. In addition, they contain the high resolution spectral radiometer data.

`sites/ua-mac/Level_1/envlog_netcdf`

- hourly files: `YYYY-MM-DD_HH-MM-SS_environmentallogger.nc`
- daily files: `envlog_netcdf_L1_ua-mac_YYYY-MM-DD.nc`

Geostreams: 5 minute observations

Data can be accessed using the geostreams API or the PEcAn meteorological workflow. These are illustrated in the sensor data tutorials.

Here is the json representation of a single five-minute observation:

Data can be accessed using the geostreams API or the PEcAn meteorological workflow.

These are illustrated in the sensor data tutorials.

Here is the json representation of a single five-minute observation from Geostreams:


```
[
   {
      "geometry":{
         "type":"Point",
         "coordinates":[
            33.0745666667,
            -111.9750833333,
            0
         ]
      },
      "start_time":"2016-08-30T00:06:24-07:00",
      "type":"Feature",
      "end_time":"2016-08-30T00:10:00-07:00",
      "properties":{
         "precipitation_rate":0.0,
         "wind_speed":1.6207870370370374,
         "surface_downwelling_shortwave_flux_in_air":0.0,
         "northward_wind":0.07488770951583902,
         "relative_humidity":26.18560185185185,
         "air_temperature":300.17606481481516,
         "eastward_wind":1.571286062845733,
         "surface_downwelling_photosynthetic_photon_flux_in_air":0.0
      }
   },
```


Variable names and units

CF standard-name	units	bety	is

air_temperature	K	airT	ta
air_temperature_max	K		ta
air_temperature_min	K		ta
air_pressure	Pa	air_pressure	
mole_fraction_of_carbon_dioxide_in_air	mol/mol		
moisture_content_of_soil_layer	kg m-2		
soil_temperature	K	soilT	
relative_humidity	%	relative_humidity	rh
specific_humidity	1	specific_humidity	N
water_vapor_saturation_deficit	Pa	VPD	
surface_downwelling_longwave_flux_in_air	W m-2	same	rl
surface_downwelling_shortwave_flux_in_air	W m-2	solar_radiation	rs
surface_downwelling_photosynthetic_photon_flux_in_air	mol m-2 s-1	PAR	
precipitation_flux	kg m-2 s-1	cccc	p
	degrees	wind_direction	
wind_speed	m/s	Wspd	
eastward_wind	m/s	eastward_wind	
northward_wind	m/s	northward_wind	

Raw Data

Data is available via Globus or Workbench:

- `/ua-mac/raw_data/co2sensor`
- `/ua-mac/raw_data/EnvironmentLogger`
- `/ua-mac/raw_data/irrigation`
- `/ua-mac/raw_data/lightning`
- `/ua-mac/raw_data/weather`

Sensor information:

- Vaisala CO2 Sensor collection
- Thies Clima Sensor collection

Computational pipeline

Environmental Logger

- **Description:** EnvironmentalLogger raw files are converted to netCDF.

Known Issues

Known issue: the irrigation data stream does not currently handle variable irrigation rates within the field. Specifically, we have not yet accounted for the Summer 2017 drought experiments. See terraref/reference-data#196 for more information.

When the full field is irrigated (as is typical), the irrigated area is 5466.1 m2 (=215.2 m x 25.4 m)

In 2017:

- Full field irrigated area from the start of the season to August 1 (103 dap) is 5466.1 m2 (=215.2 m x 25.4 m).
- Well-watered treatment zones from August 1 to 15 (103 to 116 dap): 2513.5 m2 (=215.2 m x 11.68 m) in total, combined areas of non-contiguous blocks
- Well-watered treatment zones from August 15 - 30 (116 to 131 dap): 3169.9 m2 (=215.2 m x 14.73 m), again in total as the combined areas of non-contiguous blocks

Github Issues

- computing-pipeline#252 Developmet discussions
- Downwelling Spectral radiometer calculations reference-data #30

Phenotype Data

https://terraref.github.io/tutorials/accessing-trait-data-in-r.html

Genomics data

You can access genomics data in one of the following locations:

- Download via Globus.
- The CyVerse Data Store for download or use within the CyVerse computing environment.
- The CoGe computing environment.

Please review our data use policy.

The data is structured on both the TERRA-REF strorage (accessible via Globus and Workbench) and CyVerse Data Store infrastructures as follows:

```
|-terraref
|  |-genomics
|  |  |-raw_data
|  |  |  |-bap
|  |  |  |  |-resequencing
|  |  |  |-ril
|  |  |  |  |-gbs
|  |  |-derived_data
|  |  |  |-bap
|  |  |  |  |-resequencing
|  |  |  |  |  |-danforth_center
|  |  |  |-ril
|  |  |  |  |-gbs
|  |  |  |  |  |-kansas_state
```

Whole-genome resequencing

Raw data

Raw data are in bzip2 FASTQ format, one per read pair (*_R1.fastq.bz2 and *_R2.fastq.bz2). 384 samples are available. For a list of the lines sequenced, see the sample table.

Derived data

Data derived from analysis of the raw resequencing data at the Danforth Center (version1) are available as gzipped, genotyped variant call format (gVCF) files and the final combined hapmap file.

Genotyping-by-sequencing (GBS)

Raw data

Raw data are in gzip FASTQ format. 768 samples are available. For a list of lines sequenced, see the sample table.

Derived data

Combined genotype calls are available in VCF format.

KSU Genomics Pipeline / GBS/RIL analysis

Raw Data

- genomics/raw_data/ril/gbs
  - H5JYFBCXY_1_fastq.txt
  - H5JYFBCXY_2_fastq.txt
  - Key_ril_terra

Derived Data

- genomics/derived_data/ril/gbs/kansas_state/version1/imp_TERRA_RIL_SNP.vcf

Fluorescence intensity imaging

Summary

Fluorescence intensity data is collected using the PSII camera.

Raw data access

Fluorescence intensity data is available via Clowder and Globus:

- **Clowder**: __ps2Top collection
- **Globus path**: `/sites/ua-mac/raw_data/ps2top`
- **Sensor information**: LemnaTec PSII

For details about using this data via Clowder or Globus, please see Data Access section.

Computational pipeline

Multispectral extractor

- **Description**: Raw image output is converted to a raster format (netCDF\/GeoTIFF)
- **Output**: `/sites/ua_mac/Level_1/ps2top`

Details

There are 102 bin files. The first (index 0) is an image taken right before the LED are switched on (dark reference). Frame 1 to 100 are the 100 images taken, with the LEDs on. In binary file 102 (index 101) is a list with the timestamps of each frame of the 100 frames.

Right now the LED on timespan is 1s thus the first 50 frames are taken with LEDs on the latter 50 frames with LED off..

See also

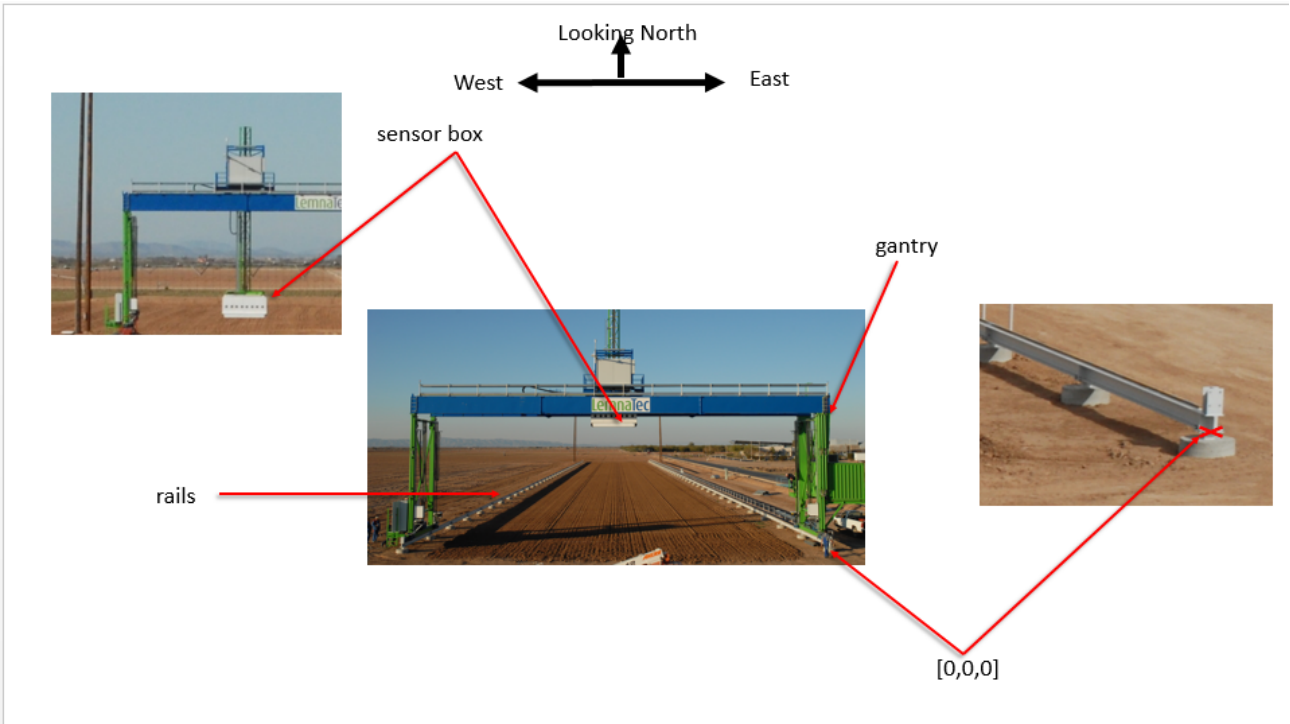
- Geospatial information

Geospatial information

Several different sensors include geospatial information in the dataset metadata describing the location of the sensor at the time of capture.

Coordinate reference systems
The Scanalyzer system itself does not have a reliable GPS unit on the sensor box. There are 3 different coordinate systems that occur in the data:

- Most common is EPSG:4326 (WGS84) USDA coordinates
- Tractor planting & sensor data is in UTM Zone 12
- Sensor position information is captured relative to the southeast corner of the Scanalyzer system in meters



EPSG:4326 coordinates for the four corners of the Scanalyzer system (bound by the rails above) are as follows:

- **NW**: 33° 04.592' N, -111° 58.505' W
- **NE**: 33° 04.591' N, -111° 58.487' W
- **SW**: 33° 04.474' N, -111° 58.505' W
- **SE**: 33° 04.470' N, -111° 58.485' W

In the trait database, this site is named the "MAC Field Scanner Field" and its bounding polygon is "POLYGON ((-111.9747967 33.0764953 358.682, -111.9747966 33.0745228 358.675, -111.9750963 33.074485715 358.62, -111.9750964 33.0764584 358.638, -111.9747967 33.0764953 358.682))"

Scanalyzer coordinates
Finally, the Scanalyzer coordinate system is right-handed - the origin is in the SE corner, X increases going from south to north, and Y increases from east to the west.

In offset meter measurements from the southeast corner of the Scanalyzer system, the extent of possible motion for the sensor box is defined as:

- **NW**: (207.3, 22.135, 5.5)
- **SE**: (3.8, 0, 0)

Scanalyzer -> EPSG:4326
1. Calculate the UTM position of known SE corner point 2. Calculate the UTM position of the target point, using SE point as reference 3. Get EPSG:4326 position based on UTM

MAC coordinates
Tractor planting data and tractor sensor data will use UTM Zone 12.

Scanalyzer -> MAC
Given a Scanalyzer(x,y), the MAC(x,y) in UTM zone 12 is calculated using the linear transformation formula:

```
ay = 3659974.971; by = 1.0002; cy = 0.0078;
ax = 409012.2032; bx = 0.009; cx = - 0.9986;
Mx = ax + bx * Gx + cx * Gy
My = ay + by * Gx + cy * Gy
```

Assume `Gx = -Gx'`, where `Gx'` is the Scanalyzer X coordinate.

MAC -> Scanalyzer

```
Gx = ( (My/cy - ay/cy) - (Mx/cx - ax/cx) ) / (by/cy - bx/cx)
Gy = ( (My/by - ay/by) - (Mx/bx - ax/bx) ) / (cy/by - cx/bx)
```

MAC -> EPSG:4326 USDA

We do a linear shifting to convert MAC coordinates in to EPSG:4326 USDA

```
Latitude: Uy = My - 0.000015258894
Longitude: Ux = Mx + 0.000020308287
```

Sensors with geospatial metadata

- stereoTop
- flirIr
- co2
- cropCircle
- PRI
- scanner3dTop
- NDVI
- PS2
- SWIR
- VNIR

Available data

All listed sensors


```
"gantry_system_variable_metadata": {
      "time": "08/17/2016 11:23:14",
      "position x [m]": "207.013",
      "position y [m]": "3.003",
      "position z [m]": "0.68",
      "speed x [m/s]": "0",
      "speed y [m/s]": "0.33",
      "speed z [m/s]": "0",
      "camera box light 1 is on": "True",
      "camera box light 2 is on": "True",
      "camera box light 3 is on": "True",
```

```
        "camera box light 4 is on": "True",
        "y end pos [m]": "22.135",
        "y set velocity [m/s]": "0.33",
        "y set acceleration [m/s^2]": "0.1",
        "y set decceleration [m/s^2]": "0.1"
      },
```

stereoTop

```
"sensor_fixed_metadata": {
      "cameras alignment": "cameras optical axis parallel to XAxis, perpen
      "optics focus setting (both)": "2.5m",
      "optics apperture setting (both)": "6.7",
      "location in gantry system": "camera box, facing ground",
      "location in camera box x [m]": "0.877",
      "location in camera box y [m]": "2.276",
      "location in camera box z [m]": "0.578",
      "field of view at 2m in X- Y- direction [m]": "[1.857 1.246]",
      "bounding Box [m]": "[1.857     1.246]",
    },
```

cropCircle

```
"sensor_fixed_metadata": {
      "location in gantry system": "camera box, facing ground",
      "location in camera box x [m]": "0.480",
      "location in camera box y [m]": "1.920",
      "location in camera box z [m]": "0.6",
    },
```

co2Sensor

```
"sensor_fixed_metadata": {
      "location in gantry system": "camera box, facing ground",
      "location in camera box x [m]": "0.35",
      "location in camera box y [m]": "2.62",
      "location in camera box z [m]": "0.7",
    },
```

flirIrCamera

```
"sensor_fixed_metadata": {
      "location in gantry system": "camera box, facing ground",
      "location in camera box x [m]": "0.877",
      "location in camera box y [m]": "1.361",
      "location in camera box z [m]": "0.520",
      "field of view x [m]": "1.496",
      "field of view y [m]": "1.105",
    },
```

ndviSensor

```
"sensor_fixed_metadata": {
      "location in gantry system": "top of gantry, facing up, camera box,
      "location in camera box x [m]": "0.33",
      "location in camera box y [m]": "2.50",
    },
```

priSensor

```
"sensor_fixed_metadata": {
      "location in gantry system": "top of gantry, facing up, camera box,
      "location in camera box x [m]": "0.400",
      "location in camera box y [m]": "2.470",
    },
```

SWIR

```
"sensor_fixed_metadata": {
      "location in gantry system": "camera box, facing ground",
      "location in camera box x [m]": "0.877",
      "location in camera box y [m]": "2.325",
```

```
        "location in camera box z [m]": "0.635",
        "field of view y [m]": "0.75",
        "optics focal length [mm]": "25",
        "optics focus apperture": "2.0",
      },
```

field scanner plots

There are 864 (54*16) plots in total and the plot layout is described in the plot plan table.

dimension	value
# rows	32
# rows / plot	2
# plots (2 rows ea)	864
# ranges	54
# columns	16
row width (m)	0.762
plot length (m)	4
row length (m)	3.5
alley length (m)	0.5

The boundary of each plot changes slightly each planting season. The scanalyzer coordinates of each plot are transformed into the (EPSG:4326) USDA coordinates using the equations above.

Hyperspectral imaging data

Summary

Hyperspectral imaging data is collected using the Headwall VNIR and SWIR sensors. In the Nov 2017 Beta Release only VNIR data is provided because we do not have the measurements of downwelling spectral radiation required by the pipeline.

Please see the README hyperspectral pipeline README for more information about how the data are generated and known issues.

Hyperspectral Algorithm and pipeline

See **Hyperspectral extractor**

Data Access

Raw Data

Raw data is available in the filesystem, accessible via Globus in the following directories:

- VNIR: `/sites/ua-mac/raw_data/VNIR`
- SWIR: `/sites/ua-mac/raw_data/SWIR`

These files are uncalibrated; see the hyperspectral pipeline repository for information on how these can be processed.

Level 1 data access

Hyperspectral data is available via Clowder, Globus #Terraref endpoint, the TERRA REF Workbench, and our THREDDS server:

- **Clowder**:
  - VNIR Hyperspectral NetCDFs
  - SWIR Collection: *Level 1 data not available*
- **Globus and Workbench**:
  - VNIR: `/sites/ua-mac/Level_1/vnir_netcdf`
  - SWIR: *Level 1 data not available*
- **Sensor information**:
  - Headwall SWIR
  - Headwall VNIR

For details about using this data via Clowder or Globus, please see Data Access section.

Level 2 data access

Level 2 data are spectral indices computed at the same resolution as Level 1. These can be found in the same Level 1 directories as their parents, but the files are appended *_ind.nc.

To get a list of hyperspectral indices currently generated https://terraref.ncsa.illinois.edu/bety/api/v1/variables?type=Reflectance%20Index

```
traits::
```

Level 3 data access

Hyperspectral Indices

The following indices are computed and provided as both Level 2 data at full spatial resolution and as Level 3 (plot level) means.

Citations can be found Morris, Geoffrey P., Davina H. Rhodes, Zachary Brenton, Punna Ramu, Vinayan Madhumal Thayil, Santosh Deshpande, C. Thomas Hash et al. "Dissecting genome-wide association signals for loss-of-function phenotypes in sorghum flavonoid pigmentation traits." G3: Genes, Genomes, Genetics 3, no. 11 (2013): 2085-2094.

Index	Label	Formula	Citation
DWSI1	Disease Water Stress Index 1	R800 / R1660	Apan, Held, Phinn and Markley (2003)
ND900_680	Normalized Difference 900/680	(R900 - R680)/(R900 + R680)	Rouse et al. (1973)
SR900_680	Simple ratio 900/680	R900/R680	Rouse et al. (1973)
DWSI2	Disease Water Stress Index 2	R1660 / R550	Apan, Held, Phinn and Markley (2003)
TCARI	Transformed Chlorophyll Absorption Ratio	3 *((R700 - R670) - 0.2* (R700 - R550) * (R700/R670))	Haboudane et al. (2002)
DWSI3	Disease Water Stress Index 3	R1660 / R680	Apan, Held, Phinn and Markley (2003)
DWSI4	Disease Water Stress Index 4	R550 / R680	Apan, Held, Phinn and Markley (2003)
DWSI5	Disease Water Stress Index 5	(R800 + R550) / (R1660 + R680)	Apan, Held, Phinn and Markley (2003)
SR700_670	Simple Ratio 700/670	R700/R670 Part of TCARI index	

RDVI	Renormalized Difference Vegetation Index	(R800 - R670)/(R800 + R670)^0.5	Rougean and Breon (1995)
PRI531	Normalized Difference 531/570 Photochemical Reflectance Index 531/570	(R531 - R570)/(R531 + R570)	Gamon et al. (1992)
EVI	Enhanced Vegetation Index	2.5 *(R800 - R680) / (R800 + 6.0f* R680 - 7.5f * R450 + 1.0f)	Huete et al. (1997)
ARVI	Atmospherically Resistant Vegetation Index	(R800 - (2.0f *R680 - R450)) / (R800 + (2.0f* R680 - R450))	Kaufman and Tanr√© (1996)
REIP1	Red-Edge Inflection Point 1	700 + 40 * {[(R670 + R780)/2 - R700] /(R740 - R700)}	Guyot and Baret, 1988
TVI	Triangular Vegetation Index	0.5 *(120* (R750 - R550) - 200 * (R670 - R550))	Haboudaneet al. (2004)
GEMI	Global Environmental Monitoring Index	((2 *(pow(R800) - pow(R680)) + 1.5* 800 + 0.5 *680) / (800 + 680 + 0.5)* (1.0 - 0.25 *(2.0f* (pow(800) - pow(680)) + 1.5 *800 + 0.5* 680) / (800 + 680 + 0.5))) - ((680 - 0.125) / (1.0 - 680))	Pinty and Verstraete (1992)
GARI	Green Atmospherically Resistant Index	(R800 - (R550 - 1.7 *(R450 - R680))) / (R800 + (R550 - 1.7* (R450 - R680)))	Gitelson et al. (1996)
DVI	Difference Vegetation Index	R800 - R680	Tucker et al. (1979)

GDVI	Green Difference Vegetation Index	R800 - R550	Sripada et al. (2006)
GNDVI	Green Normalized Difference Vegetation Index	(R800 - R550) / (R800 + R550)	Gitelson and Merzlyak (1998)
GRVI	Green Ratio Vegetation Index	R800 / R550	Sripada et al. (2006)
SR750_710	Simple Ratio 750/710 Zarco-Tejada & Miller 2001	R750/R710	Zarco-Tejada et al. (2001)
MSR705_445	Modified simple ratio 705/445	(R750 - R445)/(R705 - R445)	Sims and Gamon (2002)
WI	Water index	R900 - R970	Penuelas. et al. (1993)
Chl index	Chlorophyll index	R750/R550	Gitelson and Merzlyak (1994)
NDVI705	Normalized Difference 750/705 Chl NDI	(R750 - R705)/(R750 + R705)	Gitelson and Merzlyak (1994)
ChlDela	Chlorophyll content	(R540 - R590)/(R540 + R590)	Delaieux et al. (2014)
FRI2	Fluorescence ratio indices 2	R740/R800	Dobrowski et al. (2005)
NDVI1	Normalized Difference Vegetation Index1	(R800 - R670)/(R800 + R670)	Rouse et al. (1973)
FRI1	Fluorescence ratio index1	R690/R600	Dobrowski et al. (2005)

OSAVI	Optimized Soil Adjusted Vegetation Index	(1 + 0.16) * (R800 - R670)/(R800 + R670 + 0.16)	Rondeaux et al. (1996)
NDRE	Normalized Difference 790/720 Normalized difference red edge index	(R790 - R720)/(R790 + R720)	Barnes et al. (2000)
Car1Black	Carotenoid index from Blackburn 1998	R800/R470	Blackburn (1998)
SIPI	Structure intensive pigment index	(R800 - R450)/(R800 + R650)	Penuelas. et al. (1995)
AntGitelson	Anthocyanin (Gitelson)	(1/R550 - 1/R700) * R780	Gitelson et al. (2003,2006)
Car2Black	Carotenoid index 2 from Blackburn 1998	(R800 - R470)/(R800 + R470)	Blackburn (1998)
PRI586	Photochemical reflectance index from Panigada et al 2014	(R531 - R586)/(R531 + R586)	Panigada et al. (2014)
AntGamon	Anthocyanin from Gamon and Surfus 1999	R650/R550	Gamon and Surfus (1999)
CarChap	Carotenoid index (Chappelle)	R760/R500	Chappelle et al. (1992)
PRI512	Photochemical reflectance index from Hernandez-Clemente et al 2011	(R531- R512)/(R531 + R512)	Hern√°ndez-Clemente et al. (2011)

TCARI_OSAVI	Transformed Chlorophyll Absorption in Reflectance Index/Optimized Soil-Adjusted Vegetation Index: TCARI/OSAVI	TCARI/OSAVI	Haboudane et al. (2002)
IPVI	Infrared Percentage Vegetation Index	R800 / (R800 + R680)	Crippen et al. (1990)
NLI	Non-Linear Index	(pow(R800, 2) - R680) / (pow(R800, 2) + R680)	Goel and Qin (1994)
MNLI	Modified Non-Linear Index	((pow(R800, 2) - R680) * 1.5f) / (pow(R800, 2) + R680 + 0.5f)	Yang et al. (2008)
SAVI	Soil Adjusted Vegetation Index	(1.5f * (R800 - R680)) / (R800 + R680 + 0.5f)	Huete et al. (1988)
TDVI	Transformed Difference Vegetation Index	sqrt(0.5f + ((R800 - R680) / (R800 + R680)))	Bannari et al. (2002)
VARI	Visible Atmospherically Resistant Index	(R550 - R680) / (R550 + R680 - R450)	Gitelson et al. (2002)
RENDVI	Red Edge Normalized Difference Vegetation Index	(R750 - R705) / (R750 + R705)	Gitelson and Merzlyak (1994)
mRESR	Modified Red Edge Simple Ratio Index	(R750 - R445) / (R750 + R445)	Sims and Gamon (2002)
mRENDVI	Modified Red Edge Normalized Difference Vegetation Index	(R750 - R705) / (R750 + R705 - 2.0f * R445)	Sims and Gamon (2002)

VOG1	Vogelmann Red Edge Index 1	R740 / R720	Vogelmann et al. (1993)
VOG2	Vogelmann Red Edge Index 2	(R734 - R747) / (R715 + R726)	Vogelmann et al. (1993)
VOG3	Vogelmann Red Edge Index 3	(R734 - R747) / (R715 + R720)	Vogelmann et al. (1993)
MCARI	Modified Chlorophyll Absorption Reflectance Index	((R700 - R670) - 0.2f *(R700 - R550))* (R700 / R670)	Daughtry et al. (2000)
MCARI1	Modified Chlorophyll Absorption Reflectance Index Improved 1	1.2f *(2.5f* (R790 - R670) - 1.3f * (R790 - R550))	Haboudane et al. (2004)
MCARI2	Modified Chlorophyll Absorption Reflectance Index Improved 2	(1.5f *(2.5f* (R800 - R670) - 1.3f *(R800 - R550))) / sqrt(pow(2.0f* R800 + 1.0f, 2) - 6.0f *R800 - 5.0f* sqrt(R670) - 0.5f)	Haboudane et al. (2004)
MTVI	Modified Triangular Vegetation Index	1.2f *(1.2f* (R800 - R550) - 2.5f * (R670 - R550))	Haboudane et al. (2004)
MTVI2	Modified Triangular Vegetation Index Improved	1.5f *(1.2f* (R800 - R550) - 2.5f *(R670 - R550)) / sqrt(pow(2.0f* R800 + 1.0f ,2) - (6.0f *R800 - 5.0f* sqrt(R670)) - 0.5f)	Haboudane et al. (2004)
GMI1	Gitelson and Merzlak Index 1	R750 / R550	Gitelson and Merzlak (1997)
GMI2	Gitelson and Merzlak Index 2	R750 / R700	Gitelson and Merzlak (1997)

Lic1	Lichtenthaler Index 1	(R790 - R680) / (R790 + R680)	Lichtenthaler et al. 1996
Lic2	Lichtenthaler Index 2	R440 / R690	Lichtenthaler et al. 1996
Lic3	Lichtenthaler Index 3	R440 / R740	Lichtenthaler et al. 1996
NDNI	Normalized Difference Nitrogen Index	(log(1.0f / R1510) - log(1.0f / R1680)) / (log(1.0f / R1510) + log(1.0f / R1680))	Fourty et al. (1996)
MSR	Modified Simple Ratio	((R800 / R680) - 1.0f) / (sqrt(R800 / R680) + 1)	Chen et al. (1996)
LAI	Leaf Area Index	3.618f *((2.5.0f* (R800 - R680)) / (R800 + 6.0f *R680 - 7.5.0f* R450 + 1.0f)) - 0.118f	Boegh et al. (2002)
NRI1510	Nitrogen Related Index NRI1510	(R1510 - R660) / (R1510 + R660)	Herrmann et al. (2009)
NRI850	Nitrogen Related Index NRI850	(R850 - R660) / (R850 + R660)	Behrens et al. (2006)
NDLI	Normalized Difference Lignin Index	(log(1.0f / R1754) - log(1.0f / R1680)) / (log(1.0f / R1754) + log(1.0f / R1680))	Melillo et al. (1982)
CAI	Cellulose Absorption Index	(0.5 * (R2000 - R2200)) / R2100	Daughtry et al. (2001)
PSRI	Plant Senescence Reflectance Index	(R680 - R500) / R750	Merzlyak et al. (1999)
CRI1	Carotenoid Reflectance Index 1	1.0f / R510 - 1.0f / R550	Gitelson et al. (2002)

CRI2	Carotenoid Reflectance Index 2	1.0f / R510 - 1.0f / R700	Gitelson et al. (2002)
ARI1	Anthocyanin Reflectance Index 1	1.0f / R550 - 1.0f / R700	Gitelson et al. (2001)
ARI2	Anthocyanin Reflectance Index 2	R800 * ((1.0f / R550) - (1.0f / R700))	Gitelson et al. (2001)
SRPI	Simple Ration Pigment Index	R430 / R680	Penuelas et al. (1995)
NPQI	Normalized Phaeophytinization Index	(R415 - R435) / (R415 + R435)	Barnes et al. (1992)
NPCI	Normalized Pigment Chlorophyll Index	(R680 - R430) / (R680 + R430)	Penuelas et al. (1994)
WBI	Water Band Index	R900 / R970	Penuelas et al. (1995)
NDWI	Normalized Difference Water Index	(R857 - R1241) / (R700 + R1241)	Gao et al. (1995)
MSI	Moisture Stress Index	R819 / R1599	Hunt and Rock (1989)
NDII	Normalized Difference Infrared Index	(R857 - R1241) / (R700 + R1241)	Hardisky et al. (1983)
NMDI	Normalized Multiband Drought Index	(R819 - R1649) / (R819 + R1649)	Wang and Qu (2007)
HI	Healthy Index	((R534 - R698) / (R534 + R698)) - (R704 / 2.0f)	Mahlein et al. (2013)
CLSI	Cercospora Leaf Spot Index	((R698 - R570) / (R698 + R570)) - R734	Mahlein et al. (2013)

SBRI	Sugar Beet Rust Index	((R570 - R513) / (R570 + R513)) + (R704 / 2.0f)	Mahlein et al. (2013)
PMI	Powdery Mildew Index	((R520 - R584) / (R520 + R584)) + R724	Mahlein et al. (2013)
Crt1	Carter Index 1	R695 / R420	Carter (1994)
Crt2	Carter Index 2	R695 / R760	Carter (1996)
BIG2	Blue/Green Index	R450 / R550	Zarco-Tejada et al. (2005)
LSI	Leaf Structure Index	R1110 / R810	Maruthi Sridhar et al. (2007)
BRI	Browning Reflectance Index	((1.0f / R550) - (1.0f / R700)) / R800	Chivkunova et al. (2001)
G	Greenness Index	R554 / R677	

See also

- Sensor calibration
- Hyperspectral data pipeline
- Geospatial information

Infrared heat imaging data

Summary

Infrared heat imaging data is collected collected using the FLIR SC615 thermal sensor. These data are provided as geotiff image raster files as well as plot level means.

- Algorithms are in the flir2tif directory of the Multispectral extractor repository; see the readme for details.
- Sensor information: FLIR Thermal Camera collection

Level 1 Data Access

Filesystem (Globus and Workbench)

- `ua-mac/Level_1/ir_geotiff`

Clowder

To be created https://github.com/terraref/computing-pipeline/issues/391

Level 3 Data

Plot level summaries are named 'surface_temperature' in the trait database. In the future this name will be used for the Level 1 data as well. This name from the Climate Forecast (CF) conventions, and is used instead of 'canopy_temperature' for two reasons: First, because we do not (currently) filter soil in this pipeline. Second, because the CF definition of surface_temperature distinguishes the surface from the medium: "The surface temperature is the temperature at the interface, not the bulk temperature of the medium above or below." http://cfconventions.org/Data/cf-standard-names/48/build/cf-standard-name-table.html

---

Raw data access

Thermal imaging data is available via Clowder and Globus:

`/ua-mac/raw_data/flirIrCamera`

For details about using this data via Clowder or Globus, please see Data Access section.

Known Issues

- Data are unavailable for Season 4 (summer 2017 sorghum) and season 5 (winter 2017-2018 wheat).
  - Work to recover these data is ongoing; see terraref/reference-data#190
  - Problem description terraref/reference-data#182

See also

- Geospatial information

Multispectral imaging data

Meteorological data

Meteorological data will use Climate Forecasting 'standard names' and 'canonical units' conventions. CF is widely used in climate, meteorology, and earth sciences.

Here are some examples (note that we can change from canonical units to match the appropriate scale, e.g. "C" instead of "K"; time can use any base time and time step (e.g. `hours since 2015-01-01 00:00:00 UTC`, etc. But the time zone has to be UTC, where 12:00:00 is approx (+/- 15 min). solar noon at Greenwich.

CF standard-name	units
time	days since 1700-01-01 00:00:00 UTC
air_temperature	K
air_pressure	Pa
mole_fraction_of_carbon_dioxide_in_air	mol/mol
moisture_content_of_soil_layer	kg m-2
soil_temperature	K
relative_humidity	%
specific_humidity	1
water_vapor_saturation_deficit	Pa
surface_downwelling_longwave_flux_in_air	W m-2
surface_downwelling_shortwave_flux_in_air	W m-2
surface_downwelling_photosynthetic_photon_flux_in_air	mol m-2 s-1
precipitation_flux	kg m-2 s-1
irrigation_flux	kg m-2 s-1

irrigation_transport	kg s-1
wind_speed	m/s
eastward_wind	m/s
northward_wind	m/s

- standard_name is CF-convention standard names (except irrigation)
- units can be converted by udunits, so these can vary (e.g. the time denominator may change with time frequency of inputs)

Running The Pipeline

Before the Running

The pipepline is developed in Python, so a Python Interpreter is a must. Other than the basic Python standard librarys, the following third-party libraries are required:

- netCDF4 for Python
- numpy

Other than official CPython interpreter, Pypy is also welcomed, but please make sure that these third-party modules are correctly installed for the target interpreter. The pipeline can only works in Python 2.X versions (2.7 recommended) since numpy does not support Python 3.X versions.

Cloning from the Git:

```
git clone https://github.com/terraref/computing-pipeline.git
cd  computing-pipeline/scripts/environmental_logger
git checkout master
```

The extractor for this pipeline is developed and maintained by Max in branch "EnvironmentalLogger-extractor" under the same repository.

Get the Environmental Logger Pipeline to Work

To trigger the pipeline, use the following command:

```
python ${environmental_logger_source_path}/environmental_logger_json2netcdf.py
${input_JSON_file} ${output_netCDF_file}
```

Where:

- `${environmental_logger_source_path}` is where the three environmental_logger files are located
- `${input_JSON_file}` is where the input JSON files are located
- `${output_netCDF_file}` is where the users want the pipeline to export the product (netCDF file)

Please note that the parameter for the output file can be a path to either a directory or a file, and it is not necessarily to be existed. If the output is a path to a folder, the final product will be in this folder as a netCDF file that has the same name as the imported JSON file but with a different filename extension (`.nc` for standard netCDF file); if this path does not exist, environmental_logger pipeline will automatically make one.

Calculation

The calculation in the Environmental Logger is mainly finished by the module environmental_logger_calculation.py under the support of numpy.

Controlled Environment phenotype data

LemnaTec Scanalyzer 3D platform at the Donald Danforth Plant Science Center

Phenotype data is derived from images generated by the indoor LemnaTec Scanalyzer 3D platform at the Donald Danforth Plant Science Center using PlantCV. PlantCV is an image analysis package for plant phenotyping. PlantCV is composed of modular functions in order to be applicable to a variety of plant types and imaging systems. PlantCV contains base functions that are required to examine images from an excitation imaging fluorometer (PSII), visible spectrum camera (VIS), and near-infrared camera (NIR). PlantCV is a fully open source project: https://github.com/danforthcenter/plantcv. For more information, see:

Project website: http://plantcv.danforthcenter.org

Full documentation: http://plantcv.readthedocs.io/en/latest

Publications:

- https://doi.org/10.7717/peerj.4088
- https://doi.org/10.1016/j.molp.2015.06.005

To learn more about PlantCV, you can find examples in the terraref/tutorials repository, which is accessible on GitHub and in the TERRA REF workbench under tutorials/plantcv

- an ipython notebook demonstration of PlantCV plantcv/plantcv_jupyter_demo.ipynb.

For the TERRA-REF project, a PlantCV Clowder extractor was developed to analyze data from the Bellwether Foundation Phenotyping Facility at the Donald Danforth Plant Science Center. Resulting phenotype data is stored in BETYdb.

Computational pipeline

PlantCV extractor

- **Description**: Processes VIS/NIR images captured at several angles to generate trait metadata. The trait metadata is associated with the source images in Clowder, and uploaded to the configured BETYdb instance.
- **Output CSV**: `/sites/danforth/Level_1/<experiment name>`

Input

- Evaluation is triggered whenever a file is added to a dataset
- Following images must be found
  - 2x NIR side-view = NIR_SV_0, NIR_SV_90
  - 1x NIR top-view = NIR_TV
  - 2x VIS side-view = VIS_SV_0, VIS_SV_90
  - 1x VIS top-view = VIS_TV
- Per-image metadata in Clowder is required for BETYdb submission; this is how barcode/genotype/treatment/timestamp are determined.

Output

- Each image will have new metadata appended in Clowder including measures like height, area, perimeter, and longest_axis
- Average traits for the dataset (3 VIS or 3 NIR images) are inserted into a CSV file and added to the Clowder dataset
- If configured, the CSV will also be sent to BETYdb

Data access

Level 1

- **BETYdb**: https://terraref.ncsa.illinois.edu/bety

For details about accessing BETYdb, please see Data Access section and a tutorial on accessing phenotypes from the trait database on the TERRA REF Workbench in traits/04-danforth-indoor-phenotyping-facility.Rmd.

Raw Data

- **Clowder**: Bellwether Phenotyping Facility Space
- **Globus and Workbench**:
  - `/sites/danforth/raw_data/<experiment name>`

Point Cloud Data

Summary

3D point cloud data is collected using the Fraunhofer 3D laser scanner. Custom software installed at Maricopa converts .png output to the .ply point clouds. The .ply point clouds are converted to georeferenced .las files using the 3D point cloud extractor

Level 1 data products are provided in both .las and .ply formats.

For each scan, there are two .ply files representing two lasers, one on the left and the other on the right. These are combined in the .las files.

Sensor information

- Fraunhofer 3D scanner collection

For details about using this data via Clowder or Globus, please see Data Access section.

Data access

Data is available via Clowder, Globus, and Workbench.

- **Clowder**: Laser Scanner 3D LAS
- **Globus or Workbench File System**:
  - LAS `/sites/ua_mac/raw_data/laser3D_las`
  - PLY `/sites/ua_mac/raw_data/laser3D`

Computational pipeline

Raw sensor output (PLY) is converted to LAS format using the `ply2las` extractor

ply2las extractor

See also

- Geospatial information

Known issues

- The position of the lasers is affected by temperature. We plan to add a correction for temperature that will adjust for this effect. See terraref/reference-data#161

Maricopa UAV Data

Aerial Imagery

Description

Data sources

- sensor 1
- sensor 2
- sensor 3

Data access

Waiting on https://github.com/terraref/computing-pipeline/issues/167

Level 1 data

Level 3 data

-

Computational pipeline

See protocols: [user/protocols-UAV.md]

How to Access Data

Overview

TERRA-REF data can be accessed through many different interfaces: Globus, Clowder, BETYdb, CyVerse, and CoGe. Raw data is transfered to the primary compute pipeline using Globus Online. Data is ingested into Clowder to support exploratory analysis. The Clowder extractor system is used to transform the data and create derived data products, which are either available via Clowder or published to specialized services, such as BETYdb.

Resource	Use	Web User Interface	API*	clients
Sensor Data				
Globus	Browse directories; transfer large sensor files	globus.org #TERRAREF endpoint	docs.globus.org/api/	R, Python
Clowder	Browse and Download small Sensor Data	terraref.org/clowder	terraref.org/clowder/swaggerUI	Python
Trait Data				

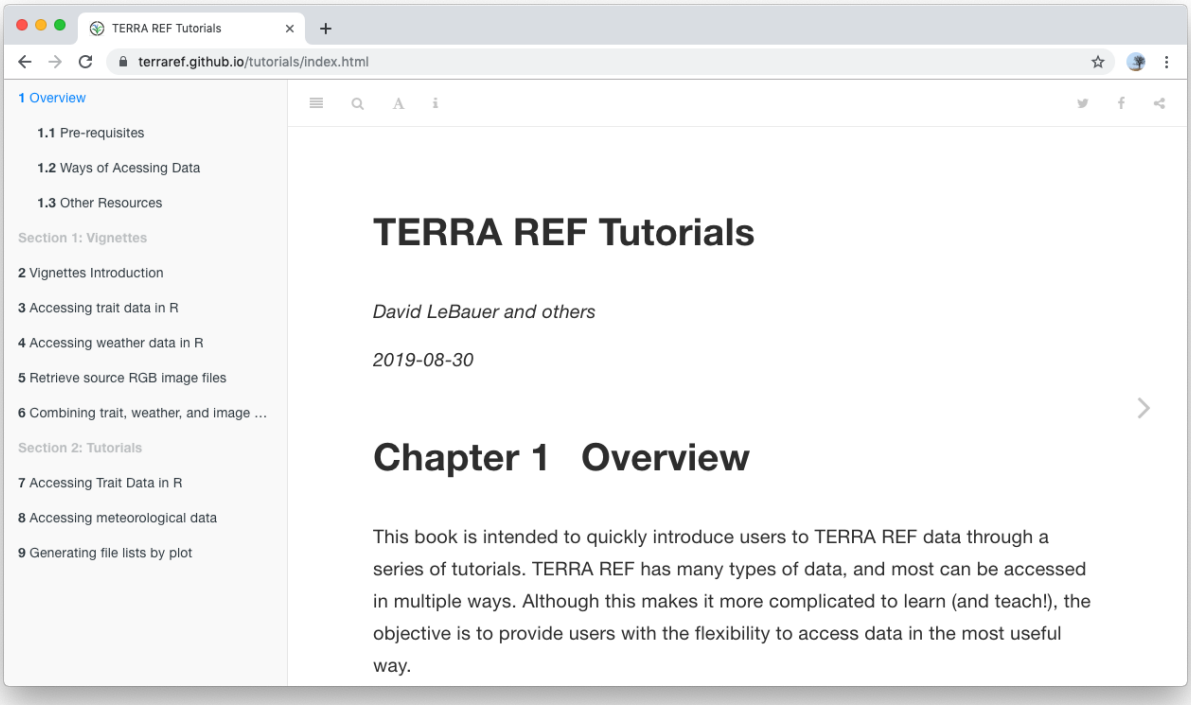
BETYdb	Trait and Agronomic Metadata	terraref.org/bety	terraref.org/bety/api/v1 and terraref.org/brapi/v1/ui	R traits package Python: terrautils SQL: Postgres in Docker
traitvis	View available trait data	terraref.org/traitvis	NA	NA
Genomics Data				
CyVerse	Download Genomics data	terraref.org/cyverse-genomics	yes	
CoGe	Download, process, visualize Genomics data	terraref.org/coge	genomevolution.org/apidocs/	
Other				
Tutorials	R and Python scripts for accessing data	terraref.org/tutorials	NA	
Advanced Search	Search across sensor and trait data	search.terraref.org (under development)	yes	

Tutorials (Recommended!)

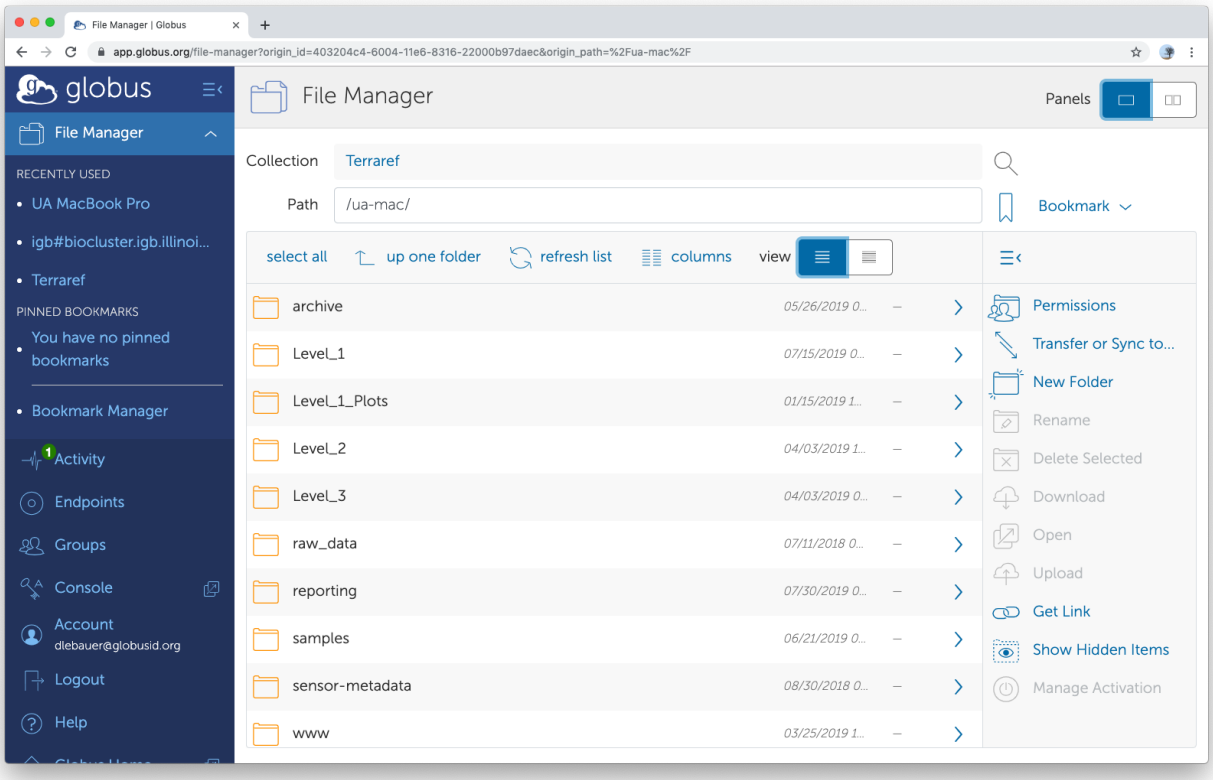
We have developed tutorials to provide users with both 'quick start' vignettes and more detailed introductions to TERRA REF datasets. Tutorials for accessing trait data, sensor data, and genomics data are organized by directory ("traits", "sensors", and "genomics").

The tutorials assume familiarity with or willingness to learn Python and / or R, and provide the greatest flexibility and access to available data.

These can be found at terraref.org/tutorials.



Globus: Browse and Transfer Files



Raw data is transferred to the primary TERRA-REF file system at the National Center for Computing Applications at the University of Illinois. Data is available for Globus transfer via the Terraref endpoint. Direct access to ROGER is restricted.

Use Globus Online when you want to transfer data from the TERRA-REF system for local analysis.

See also Globus Getting Started

Transferring data using Globus Connect:

The Globus Connect service provides high-performance, secure, file transfer and synchronization between endpoints. It also allows you to securely share your data with other Globus users.

To access data via Globus, you must first have a Globus account and endpoint.

1. Sign up for Globus at globus.org
2. Download and install Globus Connect Personal or Server.
3. Log into Globus https://www.globus.org

4. Add an endpoint for the destination (e.g. your local computer) https://www.globus.org/app/endpoints/create-gcp
5. Go to the 'transfer files' page: https://www.globus.org/app/transfer
6. Select source
    - Endpoint: #Terraref
    - Path: Navigate to the subdirectory that you want.
    - Select (click) a folder
    - Select (highlight) files that you want to download at destination
    - Select the endpoint that you set up above of your local computer or server
    - Select the destination folder (e.g. /~/Downloads/)
7. Click 'go'
8. Files will be transfered to your computer

Requesting Access to unpublished data in TERRA-REF BETYdb:

To request access to unpublished data, send your Globus id to David LeBauer (dlebauer@email.arizona.edu) with 'TERRAREF Globus Access Request' in the subject.

1. fill out the terraref.org/beta user form
2. email dlebauer@email.arizona.edu with your globusid to request access.

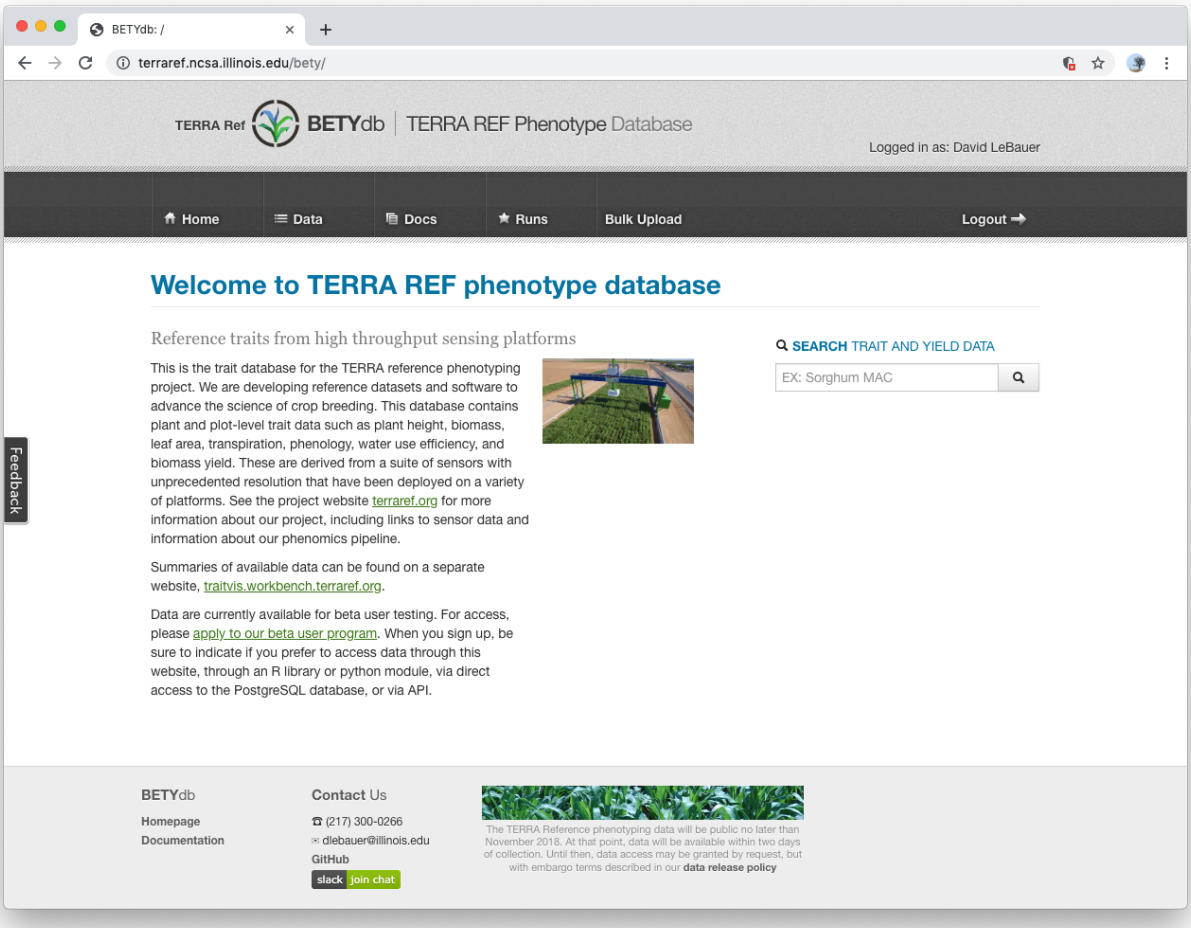
BETYdb: Trait Data and Agronomic Metadata

BETYdb is used to manage and distribute agricultural and ecological data. It contains phenotype and agronomic data including plot locations and other geolocations of interest (e.g. fields, rows, plants).

BETYdb contains the derived trait data with plot locations and other information associated with agronomic experimental design.

Accessing data in R

The easiest way to access data is to use the R traits package. This is documented in the tutorials.



Requesting Access to unpublished data in TERRA-REF BETYdb:

1. fill out the terraref.org/beta user form
2. create an account at the TERRA-REF BETYdb: terraref.org/bety (*not* betydb.org)
3. email dlebauer@email.arizona.edu for your account to be approved.

Using SQL and PostGIS with Docker (Advanced Users)

The fastest and most comprehensive way to access the database using SQL and other database interfaces (such as the R package dplyr interface described below, or GIS programs described in . You can run an instance of the database using docker, as described below

This is how you can access the TERRA REF trait database. It requires that you install the Docker software on your computer.

The easiest way to get the entire database, including metadata. Assuming you are familiar with the Postgres and / or the R dbplyr library documentation. See the TERRA REF Tutorials terraref.org/tutorials, the BETYdb Data Access guide for additional examples.

```
#git clone https://github.com/terraref/data-paper
cd data-paper/code/betydb_docker
docker-compose up -d postgres
docker-compose run --rm bety initialize
docker-compose run --rm bety sync
```

psql

```
psql -d bety -U bety -W bety
```

R

```
library(dplyr)
bety_src <- src_postgres(dbname = "bety",
                         password = 'bety',
                         host = 'localhost',
                         user = 'bety',
                         port = 5433)
```

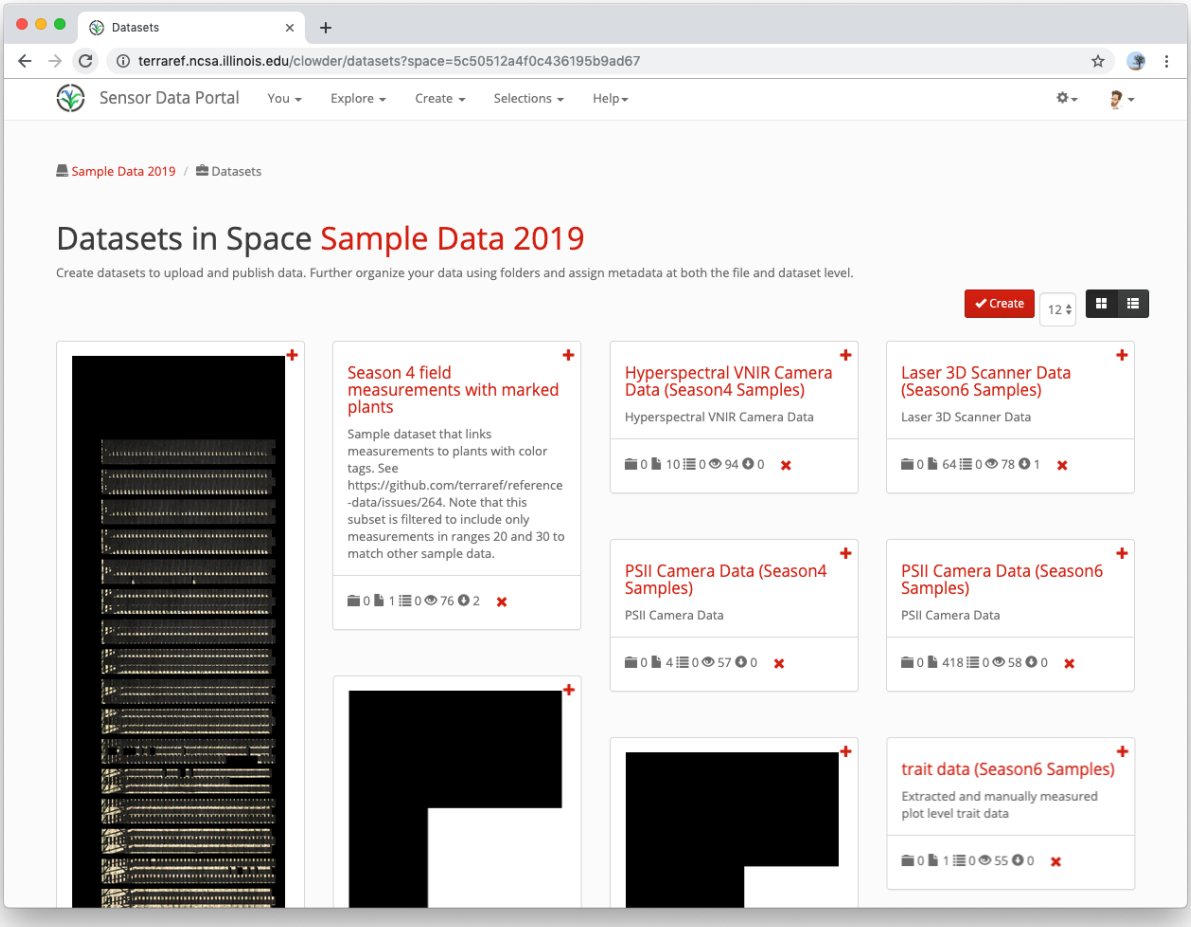
GIS software

Interested researchers can access BETYdb directly from GIS software such as ESRI ArcMap and QGIS.
In some cases direct access can simplify the use of spatial data in BETYdb. See the Appendix Accessing BETYdb with GIS Software for more information.

Clowder: Sensor Data and Metadata Browser

Clowder is an active data repository designed to enable collaboration around a set of shared datasets. TERRAREF uses Clowder to organize, annotate, and process data generated by phenotyping platforms. Datafiles are available via the Clowder web interface or API.

Clowder is the used to organize, annotate, and process raw data generated by the field scanner and other phenotyping platforms. It also stores information about sensors. Learn more about Clowder software from https://clowderframework.org



Data organization in Clowder

Data is organized into **spaces, collections,** and **datasets**, **collections**.

- **Spaces** contain collections and datasets. TERRA-REF uses one space for each of the phenotyping platforms.
- **Collections** consist of one or more datasets. TERRA-REF collections are organized by acquisition date and sensor. Users can also create their own collections.
- **Datasets** consist of one or more files with associated metadata collected by one sensor at one time point. Users can annotate, download, and use these sensor datasets.

Requesting Access to unpublished data in Clowder:

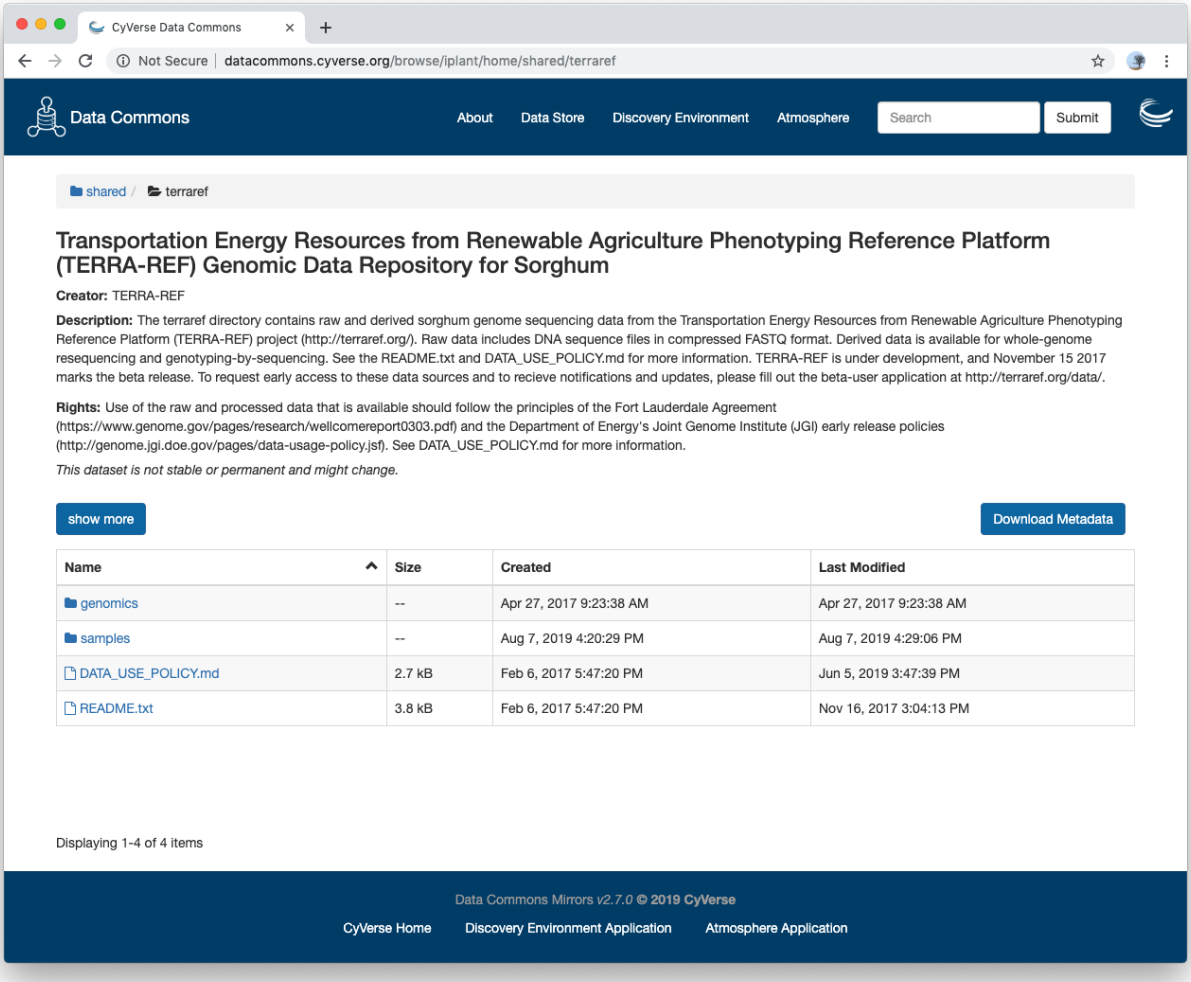
1. fill out the terraref.org/beta user form
2. create an account at the TERRA-REF Clowder site
3. email dlebauer@email.arizona.edu for your account to be approved.

CyVerse: Genomics Data

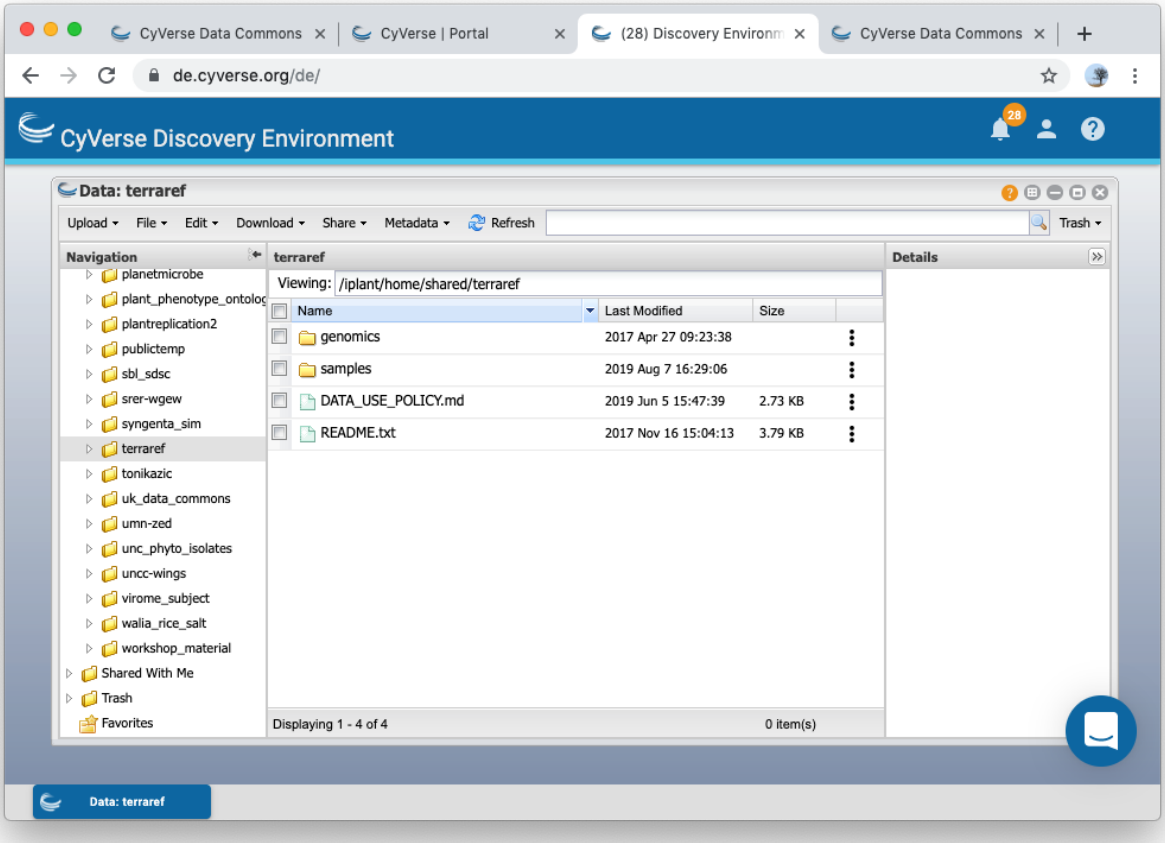
CyVerse is a National Science Foundation funded cyberinfrastructure that aims to democratize access to supercomputing capabilities.

TERRA-REF genomics data is accessible on the CyVerse Data Store and Discovery Environment. Accessing data through the CyVerse Discovery Environment requires signing up for a free CyVerse account. The Discovery Environment gives users access to software and computing resources, so this method has the advantage that TERRA-REF data can be utilized directly without the need to copy the data elsewhere.

Genomics data can be browsed and downloaded from the CyVerse data store at http://datacommons.cyverse.org/browse/iplant/home/shared/terraref



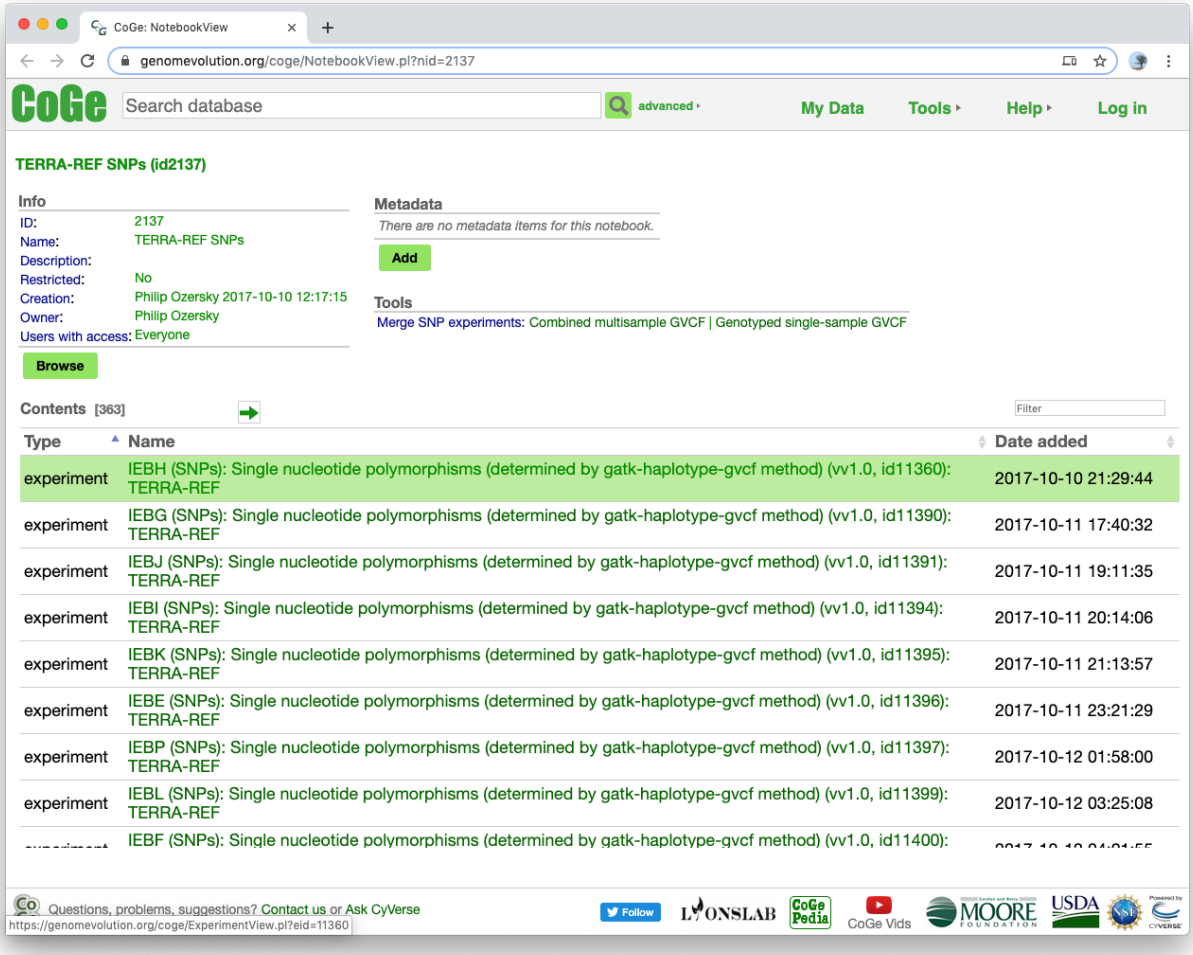
You can also find these in the CyVerse discovery environment in the TERRA-REF Community Data folder: `/iplant/home/shared/terraref`.



CoGe: Genomics Data

CoGe is a platform for performing Comparative Genomics research. It provides an open-ended network of interconnected tools to manage, analyze, and visualize next-gen data.

CoGe contains genomic information and sequence data. You can find the TERRA REF Genomics data on CoGe in this notebook:
https://genomevolution.org/coge/NotebookView.pl?nid=2137


CoGe: NotebookView
genomevolution.org/coge/NotebookView.pl?nid=2137
CoGe
Search database
advanced
My Data
Tools
Help
Log in
TERRA-REF SNPs (id2137)
Info
ID: 2137
Name: TERRA-REF SNPs
Description:
Restricted: No
Creation: Philip Ozersky 2017-10-10 12:17:15
Owner: Philip Ozersky
Users with access: Everyone
Browse
Metadata
There are no metadata items for this notebook.
Add
Tools
Merge SNP experiments: Combined multisample GVCF | Genotyped single-sample GVCF
Contents [363]
Filter
Type
Name
Date added
experiment IEBH (SNPs): Single nucleotide polymorphisms (determined by gatk-haplotype-gvcf method) (vv1.0, id11360): TERRA-REF 2017-10-10 21:29:44
experiment IEBG (SNPs): Single nucleotide polymorphisms (determined by gatk-haplotype-gvcf method) (vv1.0, id11390): TERRA-REF 2017-10-11 17:40:32
experiment IEBJ (SNPs): Single nucleotide polymorphisms (determined by gatk-haplotype-gvcf method) (vv1.0, id11391): TERRA-REF 2017-10-11 19:11:35
experiment IEBI (SNPs): Single nucleotide polymorphisms (determined by gatk-haplotype-gvcf method) (vv1.0, id11394): TERRA-REF 2017-10-11 20:14:06
experiment IEBK (SNPs): Single nucleotide polymorphisms (determined by gatk-haplotype-gvcf method) (vv1.0, id11395): TERRA-REF 2017-10-11 21:13:57
experiment IEBE (SNPs): Single nucleotide polymorphisms (determined by gatk-haplotype-gvcf method) (vv1.0, id11396): TERRA-REF 2017-10-11 23:21:29
experiment IEBP (SNPs): Single nucleotide polymorphisms (determined by gatk-haplotype-gvcf method) (vv1.0, id11397): TERRA-REF 2017-10-12 01:58:00
experiment IEBL (SNPs): Single nucleotide polymorphisms (determined by gatk-haplotype-gvcf method) (vv1.0, id11399): TERRA-REF 2017-10-12 03:25:08
IEBF (SNPs): Single nucleotide polymorphisms (determined by gatk-haplotype-gvcf method) (vv1.0, id11400):
Questions, problems, suggestions? Contact us or Ask CyVerse
https://genomevolution.org/coge/ExperimentView.pl?eid=11360
Follow
LYONSLAB
CoGe Pedia
CoGe Vids
MOORE FOUNDATION
USDA
NSF
CYVERSE

Data Use Policy

Release with Attribution

We plan to make data from the Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform (TERRA-REF) project available for use with attribution.

Please consider engaging with team members to collaborate on new research with these data. You can learn more about our approach to co-authorship and also about planned research papers in the section Manuscripts and Authorship Guidelines.

Timing and of Data Releases

We plan to release the data in stages. For access to unreleased data, please complete the beta user application to access data that has not yet been fully processed or validated.

Any access to data prior to publication is granted with the understanding that the contributions and interests of the TERRA-REF team should be recognized and respected by the users of the data. The TERRA-REF team reserves the right to analyze and published its own data. Resource users should appropriately cite the source of the data and acknowledge the resource produces. The publication of the data, as suggested in the TERRA-REF Authorship Guidelines, should specify the collaborative nature of the project, and authorship is expected to include all those TERRA-REF team members contributing significantly to the work.

The **first data publication** in Spring 2020 will provide access to data from the Maricopa field site Seasons 4 and 6 (on Dryad, citation forthcoming).

Planned **future releases** include Sorghum Season 9, data from experiments at Kansas State University, and data from the Danforth Indoor Phenotyping Facility.

Additional seasons can be requested as needed. We can provide the raw data and software required to process it. We can also collaborate with you to process the data, but this will

typically require new funding sources.

Genomic Data

Restrictions on dataset usage

Genomic data for the *Sorghum bicolor* Bioenergy Association Panel (BAP) from the TERRA-REF project is available pre-publication to maximize the community benefit of these resources. Use of the raw and processed data that is available should follow the principles of the Fort Lauderdale Agreement and the Department of Energy's Joint Genome Institute (JGI) early release policies.

By accessing these data, you agree not to publish any articles containing analyses of genes or genomic data on a whole genome or chromosome scale prior to publication by TERRA-REF and/or its collaborators of a comprehensive genome analysis ("Reserved Analyses"). "Reserved analyses" include the identification of complete (whole genome) sets of genomic features such as genes, gene families, regulatory elements, repeat structures, GC content, or any other genome feature, and whole-genome- or chromosome-scale comparisons with other species. The embargo on publication of Reserved Analyses by researchers outside of the TERRA-REF project is expected to extend until the publication of the results of the sequencing project is accepted. Scientific users are free to publish papers dealing with specific genes or small sets of genes using the sequence data. If these data are used for publication, the following acknowledgment should be included: 'These sequence data were produced by the US Department of Energy Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform (TERRA-REF) Project'. These data may be freely downloaded and used by all who respect the restrictions in the previous paragraphs. The assembly and sequence data should not be redistributed or repackaged without permission from TERRA-REF. Any redistribution of the data during the embargo period should carry this notice: "The TERRA-REF project provides these data in good faith, but makes no warranty, expressed or implied, nor assumes any legal liability or responsibility for any purpose for which the data are used. Once the sequence is moved to unreserved status, the data will be freely available for any subsequent use."

We prefer that potential users of these sequence data contact the individuals listed under Contacts with their plans to ensure that proposed usage of sequence data are not

considered Reserved Analyses.

Software and Algorithms

For algorithms, we intend to release via BSD 3 clause or MIT / BSD compatible license. Algorithms are available on GitHub in the terraref organization: github.com/terraref. Algorithms have been versioned and released on Zenodo (see https://zenodo.org/search?q=terraref).

Images, Phenotypes, and Other Raw Data

For other raw data, such as phenotypic data and associated metadata, we intend to release data under CC0: Creative Commons with No Rights Reserved - Public Domain). This follows the requirements of Dryad where the data will be published, and you can learn more about the motivation for putting data in the public domain on their blog; briefly: the CC0 license allows wide use of these data and while it does not legally bind users to acknowledge the source **data users are expected cite our data and research in publications, presentations, and other products.**

Citing Data and Software

Images, Phenotypes, Genomics, and other Data: Dryad Citation TBD

Data Processing Pipeline: Burnette, Maxwell, et al. "TERRA-REF data processing infrastructure." Proceedings of the Practice and Experience on Advanced Research Computing. 2018. 1-7. doi:10.1145/3219104.3219152

Individual Software and Documentation Components can be found on Zenodo: https://zenodo.org/search?page=1&size=20&q=terraref

Contacts

- Todd Mockler, Project/Genomics Lead (email: tmockler@danforthcenter.org)
- David LeBauer, Computing Pipeline Lead (email: dlebauer@email.arizona.edu)
- Nadia Shakoor, Project Director (email: nshakoor@danforthcenter.org)

Manuscripts and Authorship Guidelines

Summary

The willingness of many scientists to cooperate and collaborate is what makes TERRA REF possible. Because the platform encompasses a diverse group of people and relies on many data contributors to create datasets for analysis, writing scientific papers can be more challenging than with more traditional projects. We have attempted to lay out ground rules to establish a fair process for establishing authorship, and to be inclusive while not diluting the value of authorship on a manuscript. Please engage with the TERRA REF manuscript writing process knowing you are helping to forge a new model of doing collaborative scientific research.

This document is based on the Nutrient Network Authorship Guidelines, http://nutnet.org/authorship *and used with permission. Described in Borer, Elizabeth T., et al. "Finding generality in ecology: a model for globally distributed experiments."; Methods in Ecology and Evolution 5.1 (2014): 65-73.*

Copyright, Attribution, and Conditions of Use:

We plan to quickly make data and software available for use with attribution, under CC-By 4.0, MIT compatable license, or Ft. Lauderdale Agreement as described in our Data Use Guidelines. Such data can be used with attribution (e.g. citation); co-authorship opportunities are welcome where warranted (see below) by specific contributions to the manuscript (e.g. help in interpreting data beyond technical support).

We are making data available early for users under the condition that manuscripts led within the team not be scooped. In these cases, people who wish to use the data for publication prior to official open release date of November 2018 should coordinate co-authorship with the person responsible for collecting the data.

Overview of the TERRA REF authorship process:

Inclusive but not gratuitous

Our primary goals in the TERRA REF authorship process are to consistently, accurately and transparently attribute the contribution of each author on the paper, to encourage participation in manuscripts by interested scientists, and to ensure that each author has made sufficient contribution to the paper to warrant authorship.

Steps:

1. **Read these authorship policies** and guidelines.
2. **Consult the current list of manuscripts** (http://terraref.org/manuscripts) for current proposals and active manuscripts, contact the listed lead author on any similar proposal to minimize overlap, or to join forces. Also carefully read these guidelines.
3. **Prepare a manuscript proposal**. Your proposal will list the lead author(s), the title and abstract body, and the specific data types that you will use. You can also specify more detail about response and predictor variables (if appropriate), and indicate a timeline for analysis and writing. Submit your proposal through this form.
   Proposed ideas are reviewed by the authorship committee primarily to facilitate appropriate collaborations, identify potential duplication of effort, and to support the scientists who generate data while allowing the broader research community access to data as quickly and openly as possible. The authorship committee may suggest altering or combining analyses and papers to resolve issues of overlap.
4. **Circulate your draft analysis and manuscript to solicit Opt-In authorship**.
   For global analyses, the lead author should circulate the manuscript to the Network by submitting a email to the TERRA REF team.
   For analyses of more limited scope, the lead author should circulate the manuscript to network collaborators who have indicated interest at the abstract stage, those who have contributed data, and any others who the lead author deems appropriate.
   In both cases, the subject line of the email should include the phrase "OPT-IN PAPER";
   This email should also include a *deadline* by which time co-authors should respond.
   The right point to share your working draft and solicit co-authors is different for each manuscript, but in general:
   1. sharing early drafts or figures allows for more effective co-author contribution. While ideally this would mean circulating the manuscript at a very early stage for opt-in to the entire network, it is acceptable and even typical to share early drafts or figures among a smaller group of core authors.

2. circulating essentially complete manuscripts does not allow the opportunity for meaningful contribution from co-authors, and is discouraged.

5. **Potential co-authors** should signal their intention to opt-in by responding by email to the lead author before the stated deadline.
6. **Potential co-authors** should inform the lead author of any additional candidates for co-authorship who should be considered.
7. **Lead authors should** is responsible for making sure that any who have made contributions warranting co-authorship have actively opted in or out (authors should not be excluded due to a missed email or a misunderstanding of the scope of the manuscript and their contributions). The goal is to ensure that the author list is inclusive and consistent.
8. **Lead authors should** keep an email list of co-authors and **communicate regularly** about progress including sharing drafts of analyses, figures, and text as often as is productive and practical.
9. **Lead authors should** circulate complete drafts among co-authors and consider comments and changes. Given the wide variety of ideas and suggestions provided on each TERRA REF paper, co-authors should recognize the final decisions belong to the lead author.
10. **Final manuscripts should be reviewed and approved by each co-author before submission.**
11. **All authors and co-authors should** fill out their contribution in the authorship rubric and attach it as supplementary material to any TERRA REF manuscript. Lead authors are responsible for ensuring consistency in credit given for contributions, and may alter co-author's entries in the table to do so.
    The authorship rubric provides a framework for the opt-in process. Lead authors should copy the template and edit the contents for a specific manuscript, then circulate to potential co-authors.
    Note that the last author position may be appropriate to assign in some cases. For example, this would be appropriate for advisors of lead authors who are graduate students or postdocs and for papers that two people worked very closely to produce.
12. **The lead author should** carefully review the authorship contribution table to ensure that all authors have contributed at a level that warrants authorship and that contributions are consistently attributed among authors. The lead author should also ensure that all contributions that warrant co-authorship.
    - Has each author made contributions in at least two areas in the authorship rubric?
    - Did each author provide thoughtful, detailed feedback on the manuscript?
    - Have all qualified contributors actively opted in or out of co-authorship?

Authors are encouraged to contact the TERRA REF PI (Mockler) or authorship committee (Jeff White, Geoff Morris, Todd Mockler, David LeBauer, Wasit Wulamu, Nadia Shakoor) about any confusion or conflicts.

Co-authorship

Authorship must be earned through a *substantial contribution.* Traditionally, project initiation and framing, data analysis and interpretation, software or algorithm development, and manuscript preparation are all authorship-worthy contributions, and remain so for TERRA REF manuscripts. However, TERRA REF collaborators have also agreed that collaborators who lead a site from which data are being used in a paper can also opt-in as co-authors, under the following conditions: **(1)** the collaborators' site has contributed data being used in the paper's analysis; and **(2)** that this collaborator makes additional contributions to the particular manuscript, including data analysis, writing, or editing. For co-authorship on opt-out papers, each individual must be able to check **at least two** boxes in the rubric in addition to contribution to the writing process. *These guidelines apply equally to manuscripts led by graduate students.*

Co-author Expectations

This section is derived from the International Committee of Medical Journal Editors (ICMJE) Uniform Requirements for Manuscripts Submitted to Biomedical Journals.

Each author is expected to meet all of the following conditions:

- Substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data, and
- Drafting the article *or* revising it critically for important intellectual content, and
- Final approval of the version to be published, and
- Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Types of Author Contributions

Manuscripts published by TERRA REF will be accompanied by a supplemental table indicating authorship contributions. You can copy and share the authorship rubric. For opt-in papers, a co-author usually should contribute to writing and revision as well as at least two of the following areas checked in the authorship rubric. This follows the CRediT Taxonomy as published in the PLOS ONE authorship guidelines.

Contributor Role	Role Definition
Conceptualization	Ideas; formulation or evolution of overarching research goals and aims.
Data Curation	Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse.
Formal Analysis	Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data.
Funding Acquisition	Acquisition of the financial support for the project leading to this publication.
Investigation	Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection.
Methodology	Development or design of methodology; creation of models
Project Administration	Management and coordination responsibility for the research activity planning and execution.
Resources	Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools.
Software	Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.

Supervision	Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team.
Validation	Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs.
Visualization	Preparation, creation and/or presentation of the published work, specifically visualization/data presentation.
Writing – Original Draft Preparation	Creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation).
Writing – Review & Editing	Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages.

Author Order

By default, we will follow the conventions of the scientific community that is the target audience of the journal in which the article is published. This should typically follow:

- First is lead author
- Last is the supervisor of the lead author.
- if > 1 lead or senior authors these will be listed first and last, respectively, and identified in the author contributions section of the acknowledgements.
- All other contributors are listed alphabetically.

Publications committee

Members: David LeBauer, Todd Mockler, Geoff Morris, Duke Pauli, Nadia Shakoor, Wasit Wulamu

The publications committee ensures communication across projects to avoid overlap of manuscripts, works to provide guidance on procedures and authorship guidelines, and serves as the body of last resort for resolution of authorship disputes within the Network.

Acknowledgments

Please use the following text in the acknowledgments of TERRA REF manuscripts:

> The [information / data / work] presented here is from the TERRA REF experiment, funded by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000594. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Keywords

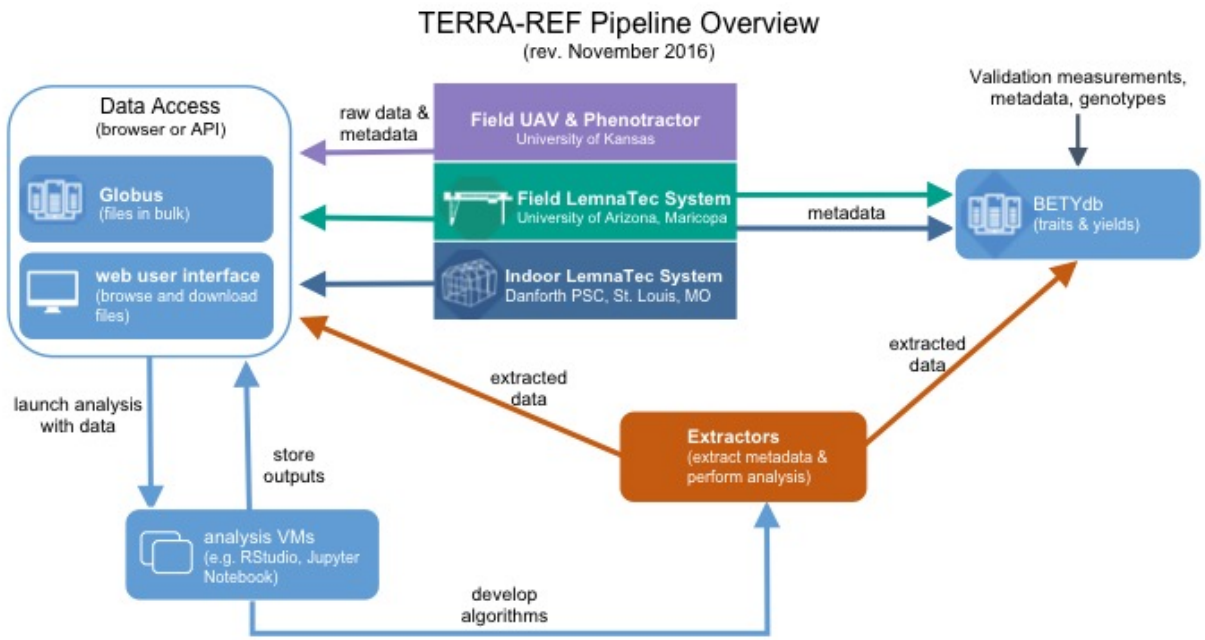
Please use "TERRA REF"; as one of your keywords on submitted manuscripts, so that TERRA REF work is easily indexed and searchable.

Technical Documentation

This section includes the following:

Software

TERRA-REF uses a suite of databases and software components to automate the analysis of sensor data to produce plant and plot level traits / phenotypes, store and provide access to data..



Databases

Clowder (sensor data and computation management with web user interface)

Clowder is the primary system used to organize, annotate, and process raw data generated by the phenotyping platforms as well as information about sensors. Use Clowder to explore the raw TERRA-REF data, perform exploratory analysis, and develop custom extractors. For more information, see ~~Using Clowder~~.

Globus Connect (large data transfer)

Raw data is transferred to the primary TERRA-REF compute pipeline using Globus Online. Globus also provides access to TERRA REF files, but this is not a primary portal and metadata in Clowder may be required to locate and interpret these files. Use Globus Online when you want to transfer data from the TERRA-REF system for local analysis by accessing the Terraref endpoint. For more information, see ~~Using Globus~~.

BETYdb (phenotype data)

BETYdb is a database and web interface to the trait / phenotype data and agronomic metadata. This is where you can find plant and plot level trait data as well as plot locations and other information associated with agronomic experimental design. Use BETYdb to access derived trait and agronomic data. For more information, see ~~Using BETYdb~~.

Algorithms

Extractors

Each step in the pipeline is performed by an algorithm. These are maintained in the TERRA REF GitHub organization in repositories with names that begin in `extractors-*` such as github.com/terraref/extractors-hyperspectral.

Plant CV

Plant CV is an imaging processing package specific for plants that is built upon open-source software platforms OpenCV, NumPy, and MatPlotLib. Plant CV is used for trait identification, the output is stored in both Clowder and BETYdb.

CoGe

CoGe is a platform for performing Comparative Genomics research. It provides an open-ended network of interconnected tools to manage, analyze, and visualize next-gen data. CoGe contains genomic information and sequence data from the TERRA REF project.

Data Standards

Overview

TERRA's data standards facilitate the exchange of genomic and phenomic data across teams and external researchers. Applying common standards makes it easier to exchange analytical methods and data across domains and to leverage existing tools.

When practical, existing conventions and standards have been used to create data standards. Spatial data adopts Federal Geographic Data Committee (FGDC) and Open Geospatial Consortium (OGC) data and meta-data standards. CF variable naming convention was adopted for meteorological data and biophysical data. Data formats and variable naming conventions were adapted from NEON and NASA.

Feedback from data creators and users were used to define the types of data formats, semantics, and interfaces, file formats, and representations of space, time, and genetic identity based on existing standards, commonly used file formats, and user needs.

We anticipate that standards and data formats will evolve over time as we clarify use cases, develop new sensors and analytical pipelines, and build tools for data format conversion and feature extraction and tracking provenance. Each year we will re-convene to assess our standards based on user needs. The Standards Committee will assess the trade-off between the upfront cost of adoption with the long-term value of the data products, algorithms, and tools that will be developed as part of the TERRA program. The specifications for these data products will be developed iteratively over the course of the project in coordination with TERRA funded projects. The focus will be to take advantage of existing tools based on these standards, and to develop data translation interfaces where necessary.

See also

- Agronomic and Phenotype Data Standards
- Environmental Data Standards
- Genomic Data Standards

- Sensor Data Standards
- Data Standards Committee

Existing Data Standards

This page summarizes existing standards, conventions, controlled vocabularies, and ontologies used for the representation of crop physiological traits, agronomic metadata, sensor output, genomics, and other inforamtion related to the TERRA-REF project.

Metadata standards

International Consortium for Agricultural Systems Applications (ICASA)

The ICASA Version 2.0 data standard defines an abstract model and data dictionary for the representation of agricultural field expirements. ICASA is explicitly designed to support implementations in a variety of formats, including plain text, spreadsheets or structured formats. It is important to note that ICASA is both the data dictionary and a format used to describe experiments.

The Agricultural Model Intercomparison Project (AgMIP) project has developed a JSON-based format for use with the AgMIP Crop Experiment (ACE) database and API.

Currently, the ICASA data dictionary is represented as a Google Spreadsheet and is not suitable for linked-data applications. The next step is to render ICASA in RDF for the TERRA-REF project. This will allow TERRA-REF to produce data that leverages the ICASA vocabulary as well as other external or custom vocabularies in a single metadata format.

The ICASA data dictionary is also being mapped to various ontologies as part of the Agronomy Ontology project. With this, it may be possible in the future to represent ICASA concepts using formal ontologies or to create mappings/crosswalks between them.

See also:

- White et al (2013). Integrated Description of Agricultural Field Experiments and Production: The ICASA Version 2.0 Data Standards. Computers and Electronics in Agriculture.
- AgMIP JSON Data Objects format description

- ICASA Master Variable List
  </small>

Minimum Information About a Plant Phenotyping Experiment (MIAPPE)

MIAPPE was developed by members of the European Phenotyping Network (EPPN) and the EU-funded transPLANT project. It is intended to define a list of attributes necessary to fully describe a phenotyping experiment.

The MIAPPE standard is available from the transPlant standards portal and is compatible with the ISA-Tools suite framework. The transPLANT standards portal also provides example configuration for the ISA toolset.

Section	Recommended ontologies
General metadata	Ongtology for Biomedical Investigations (OBI), Crop Research Ontology (CRO)
Timing and location	OBI, Gazetteer (GAZ)
Biosource	UNIPROT taxonomy, NCBI taxonomy
Environment, treatments	XEO Environment Ontology, Ontology of Environmental Features (ENVO), CRO
Experimental design	OBI, CRO, Statistics Ontology (STATO)
Observed values	Trait Ontology (TO), Plant Ontology (PO), Crop Ontology (CO), Phenotypic Quality Ontology (PATO), XEO/XEML

MIAPPE is currently the only standard listed in biosharing.org for the phenotyping domain. While several databases claim to support MIAPPE, the standard is still nascent.

MIAPPE is based on the ISA framework, building on earlier “minimum information” standards, such as MIAME (Minimum Information about a Microarray Experiment). If the MIAPPE standard is determined to be useful for TERRA-REF, it would be worth reviewing the

MIAME steandard and related formats such as MAGE-TAG, MINiML, and SOFT accepted by the Gene Expression Omnibus (GEO). GEO is a long-standing repository for genetic research data and might serve as another model for TERRA-REF.

It is worth noting that linked-data methods are supported but optional when depositing data to GEO. The MAGE-TAB format, similar to the MIAPPE ISA Tab format, does support sources for controlled vocabulary terms or ontologies.

See also:

- Minimum Information about a Plant Phenotyping Experiment
  </small>

Dublin Core Application Profiles

While some communities define explicit metadata schema (e.g., Ecological Metadata Language), another approach is the use of "application profiles." An application profile is declaration of metadata terms adopted by a community or an organization along with the source of the terms. Application profiles are composed of terms drawn from multiple vocubularies or ontologies to define a "schema" or "profile" for metadata. For example, the Dryad metadata profile draws on the Dublin Core, Darwin Core, and Dryad-specific elements.

See also:

- DCMI Guidelines for Dublin Core Application Profiles.
- Example Dryad Metadata Profile
- DCMI Singapore Framework
  </small>

Trait Dictionary Format (Crop Ontology)

The Crop Ontology curation tool supports import and export of trait information in a trait dictionary format.

See also:

- The Crop Ontology Improving the Quality of 18 Crop Trait Dictionaries
  </small>

Vocabularies and Ontologies

This section reviews related controlled vocabularies, data dictionaries, and ontologies.

Biofuel Ecophysiological Traits and Yields Database (BETYdb)

While BETYdb is not a controlled vocabulary itself, the relational schema models a variety of concepts including managements, sites, treatments, traites, and yields.

The BETYdb “variables” table defines variables used to represent traits in the BETYdb relational model. There has been some effort to standardize variable names by adopting Climate Forecasting (CF) convention standard names where variables overlap. A variable is represented as a name, description, units, as well as min/max values.

For example:


```
"variable": {
    "created_at": "2016-03-07T11:23:58-06:00",
    "description": "",
    "id": 604,
    "label": "",
    "max": "1000",
    "min": "0",
    "name": "NDVI",
    "notes": "",
    "standard_name": "normalized_difference_vegetation_index",
    "standard_units": "ratio",
    "type": "",
    "units": "ratio",
    "updated_at": "2016-03-07T11:26:07-06:00"
}
```


See also:

- The full suite of variables supported by BETYdb
- Trait variables used in the TERRA Ref BETYdb instance
  </small>

DCMI Metadata terms

Controlled vocabulary for the representation of bibliographic information. *See also:*

- DCMI Terms
  </small>

Climate and Forecast Standard Name Table

Standard variable names and naming convention for use with NetCDF. The Climate and Forecast metadata conventions are intended to promote sharing of NetCDF files. The CF conventions define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities.

Basic conventions include lower-case letters, numbers, underscores, and US spelling.

Information is encoded in the variable name itself. The basic format is (optional components in []):

```
[surface] [component] standard_name [at surface] [in medium] [due to process]
[assuming condition]
```

For example:

Standard names have optional canonical units, AMIP and GRIB (GRidded Binary) codes.

The CF standard names have been converted to RDF by several communities, including the Marine Metadata Interoperability (MMI) project.

Dimensions: time, lat, lon, other specify time first (unlimited) lat, lon or x, y extent to field boundaries.

See also:

- CF Conventions
- CF Conventions FAQ mentions RDF conversions.
  </small>

ICASA master variable list

Vocabulary and naming conventions for agricultural modeling variables, used by AgMIP. The ICASA master variable list is included, at least in part, in the AgrO ontology. The NARDN-HD Core Harmonized Crop Experiment Data is also taken from the ICASA vocabulary.

ICASA variables have a number of fields, including name, description, type, min and max values.

See also:

- ICASA Master Variable List
- White et al (2013). Integrated Description of Agricultural Field Experiments and Production: The ICASA Version 2.0 Data Standards. Computers and Electronics in Agriculture.
  </small>

NARDN-HD Core Harmonized Crop Experiment Data

A subset of the ICASA data dictionary representing set of core variables that are commonly collected in field crop experiments. These will be used to harmonize data from USDA experiments as part of a National Agricultural Research Data Network.

CSDMS Standard Names

Variable naming rules and patterns for any domain developed as part of the CSDMS project as an alternative to CF. CSDMS standard names is considered to have a more flexible community approval mechanism than CF. CSDMS names include object, quantity/attribute parts.

CSDMS names have been converted to RDF as part of the Earth Cube Geosemantic Server project.

See also:

- CSMDS Standard Names
  </small>

International Plant Names Index (IPNI)

http://www.ipni.org/

IPNI is a database of the names and associated basic bibliographical details of seed plants, ferns and lycophytes. It's goal is to eliminate the need for repeated reference to primary sources for basic bibliographic information about plant names.

NCBI Taxonomy

http://www.ncbi.nlm.nih.gov/taxonomy

A curated classification and nomenclature for all of the organisms in the public sequence databases that represents about 10% of the described species of life on the planet. Taxonomy recommended by MIAPPE.

Ontologies

Agronomy Ontology (AGRO)

The Agronomy Ontology “describes agronomic practices, agronomic techniques, and agronomic variables used in agronomic experiments.” It is intended as a complementary ontology to the Crop Ontology (CO). Variables are selected out of the International Consortium for Agricultural Systems Applications (ICASA) vocabulary and a mapping between AgrO and ICASA is in progress. AgrO is intended to work with the existing ontologies including ENVO, UO, PATO, IAO, and CHEBI. It will be part of an Agronomy Management System and fieldbook modeled on the CGIAR Breeding Management System to capture agronomic data.

See also:

- OBO Foundry. Agonomy Ontology
- FAO. Crop Ontology: harmonizing semantics for phenotyping and agronomy data
- RDA. Interest Group on Agricultural Data (IGAD)
  </small>

Crop Ontology (CO)

The Crop Ontology (CO) contains "Validated concepts along with their inter-relationships on anatomy, structure and phenotype of crops, on trait measurement and methods as well as on Germplasm with the multi-crop passport terms." The ontology is actively used by the CGIAR community and a central part of the Breeding Management System. MIAPPE recommends the CO (along with TO, PO, PATO, XEML) for observed variables.

Shrestha et al (2012) describe a method for representing trait data via the CO.

See also:

- Crop Ontology
- Shrestha et al (2012). Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. Front Physiol. 2012 Aug 25;3:326.
  </small>

Crop Research Ontology (CRO)

Describes experimental design, environmental conditions and methods associated with the crop study/experiment/trial and their evaluation. CRO is part of the Crop Ontology platform, originally developed for the International Crop Information System (ICIS). CRO is recommended in the MIAPPE standard for general metadata, environment, treatments, and experimental design fields.

See also:

- Crop Research Ontology
- International Crop Information System
  </small>

Extensible Observation Ontology (OBOE)

Cited in Kattge et al (2011) as an example of an ontology used in ecology and environmental sciences to represent measurements and observation. However, the CRO may be better suited for TERRA-REF.

See also:

- Kattge, J.(2011). A generic structure for plant trait databases
  </small>

Gene Ontology (GO)

Defines concepts/classes used to describe gene function, and relationships between these concepts. GO is a widely-adopted ontology in genetics research, supported by databases such as GEO. This ontology is cited in Krajewski et al (2015) and might be relevant for the TERRA genomics pipeline.

See also:

- Gene Ontology
- Krajewski et al (2015). Towards recommendations for metadata and data handling in plant phenotyping. Journal of Experimental Botany, 66(18), 5417–5427.
  </small>

Information Artifact Ontology (IAO)

Information entities, originally driven by work by OBI (e.g., abstract, author, citation, document etc). IAO covers similar territory to the Dublin Core vocabulary.

Ontology for Biomedical Investigations (OBI)

Integrated ontology for the description of biological and clinical investigations. This includes a set of 'universal' terms, that are applicable across various biological and technological domains, and domain-specific terms relevant only to a given domain. Recommended by MIAPPE for general metadata, timing and location, and experimental design.

See also:

- Minimum Information about a Plant Phenotyping Experiment
  </small>

Phenotype and Attribute Ontology (PATO)

Phenotypic qualities (properties).

Recommended in MAIPPE for use in the observed values field.

See also:

- Minimum Information about a Plant Phenotyping Experiment
  </small>

Plant Environment Ontology (EO)

Part of the Plant Ontology (PO), standardized controlled vocabularies to describe various types of treatments given to an individual plant / a population or a cultured tissue and/or cell type sample to evaluate the response on its exposure.

Plant Ontology (PO)

Describes plant anatomy and morphology and stages of development for all plants intended to create a framework for meaningful cross-species queries across gene expression and phenotype data sets from plant genomics and genetics experiment. Recommended by MIAPPE for observed values fields. Along with EO, GO, and TO make up the Gramene database. Links plant anatomy, morphology and growth and development to plant genomics data.

See also:

- Minimum Information about a Plant Phenotyping Experiment
  </small>

Plant Trait Ontology (TO)

Along with EO, GO, and PO, make up the Gramene database to link plant anatomy, morphology and growth and development to plant genomics data. Recommended by MIAPPE for observed values fields.

Example trait entry:

```
[Term]
id: TO:0000019
name: seedling height
def: "Average height measurements of 10 seedlings, in centimeters from the
synonym: "SH" RELATED []
is_a: TO:0000207 ! plant height
```

See also:

- Minimum Information about a Plant Phenotyping Experiment
  </small>

Statistics Ontology (STATO)

General purpose statistics ontology coveraging processes such as statistical tests, their conditions of application, and information needed or resulting from statistical methods,

such as probability distributions, variables, spread and variation metrics. Recommended by MIAPPE for experimental design.

See also:

- Minimum Information about a Plant Phenotyping Experiment
  </small>

Units of Measurement Ontology (UO)

Metric units for PATO. This OBO ontology defines a set of prefixes (giga, hecto, kilo, etc) and units (area/square meter, volume/liter, rate/count per second, temperature/degree Fahrenheit). The two top-level classes are prefixes and units.

UO is mentioned in relation to the Agronomy Ontology (AGRO), but PATO is also recommended by MIAPPE for observed values fields

While there are general standard units, it seems unlikely that these would ever be gathered in a single place. It seems more useful to define a high-level ontology to represent a "unit" and allow domains and communities to publish their own authoritative lists.

XEML Environment Ontology (XEO)

Created to help plant scientists in documenting and sharing metadata describing the abiotic environment.

DDI-RDF Discovery Vocabulary

Data Catalog Vocabulary (DCAT)

The Data Catalog Vocabulary is an RDF vocabulary intended to facilitate interoperability between data catalogs published on the Web. DCAT defines a set of classes including Dataset, Catalog, CatalogRecord, and Distribution.

Data Cite Ontology

The DataCite Ontology

Data Cube Vocabulary

The Data Cube Vocabulary is an RDF-based model for publishing multi-dimentional datasets, based in part on the SDMX guidelines. DataCube defines a set of classes including DataSet, Observation, and MeasureProperty that may be relevant to the TERRA project.

Statistical Data and Metadata Exchange (SDMX)

SDMX is an international initiative for the standarization of the exchange of statistical data and metadata among international organizations. Sponsors of the initiative include Eurostat, European Central Bank, the OECD, World Bank and the UN Statistical Division. They have defined a framework and an exchange format, SDMX-ML, for data exchange. Community members have also developed RDF encodings of the SDMX guidelines that are heavily referenced in the Data Cube vocabulary examples.

Related Software, Services, and Databases

Standard formats, ontologies, and controlled vocabularies are typically used in the context of specific software systems.

Agricultural Model Inter-Comparison and Improvement Project (AgMIP) Crop Experiment (ACE) Database

AgMIP "seeks to improve the capability of ecophysiological and economic models to describe the potential impacts of climate change on agricultural systems.
AgMIP protocols emphasize the use of multiple models; consequently, data harmonization is essential. This interoperability was achieved by establishing a data exchange mechanism with variables defined in accordance with international standards; implementing a flexibly structured data schema to store experimental data; and designing a method to fill gaps in model-required input data."

The data exchange format is based on a JSON rendering of the ICASA Master Variable List. Data are transfer into and out of the AgMIP Crop Experiment (ACE) and AgMIP Crop Model (ACMO) databases via REST apis using these JSON objects.

See also

- AgMIP Crop Expirement Database
- Porter et al (2014). Harmonization and translation of crop modeling data to ensure interoperability. Environmental Modelling and Software. 62:495-508.
- AgMIP Data Products presentation
- AgMIP on Github
- AgMIP Crop Experiment Database data variables
- AgMIP API
- AgMIP using ICASA standards</small>

Biofuel Ecophysiological Traits and Yields Database (BETYdb)

BETYdb is used to store TERRA meta-data, provenance, and traits information.

BETYdb traits are available as web-page, csv, json, xml. This can be extended to allow spatial, temporal, and taxonomic / genomic queries. Trait vectors can be queries and rendered in several output formats. For example:

Here are some examples from betydb.org.

- HTML output
- csv output
- xml output
- Json-compatible output

A separate instance of BETYdb is maintained for use by TERRA Ref at terraref.ncsa.illinois.edu.org/bety. The scope of the TERRA Ref database is limited to high througput phenotyping data and metadata produced and used by the TERRA program. Users can set up their own instances of BETYdb and import any public data in the distributed BETYdb network.

See also: BETYdb documentation

- BETYdb Data Access includes accessing data with web interface, API, and R traits package
- BETYdb constraints, see section "uniqueness constraints"
- BETYdb Data Entry
  </small>

Gramene

Gramene is a curated, open-source, integrated data resource for comparative functional genomics in crops and model plant species

Integrated Breeding Platform/Breeding Management System

System for managing the breeding process including lists of germplasms, defining crosses, managing nurseries, trials, as well as ontologies and statistical analysis.

See also:

- BMS Site
  </small>

TERRA Ref has an instance of BMS hosted by CyVerse (requires login).

International Crop Information System

ICIS is "a database system that provides integrated management of global information on crop improvement and management both for individual crops and for farming systems." ICIS is developed by Consultative Group for International Agricultural Research (CGIAR).

See also

- Fox and Skovmand (1996). "The International Crop Information System (ICIS) - connects genebank to breeder to farmer's field." Plant adaptation and crop improvement, CAB International.
  </small>

MODAPS NASA MODIS Satellite data

The MODAPS NASA MODIS Satellite data encompasses a library of functions that provides programmatic data access and processing services to MODIS Level 1 and Atmosphere data products. These routines enable both SOAP and REST based web service calls against the data archives maintained by MODAPS. These routines mirror existing LAADS Web services.

See also:

- NDISC Modis Data Summaries
  </small>

Phenomics Ontology Driven Database (PODD)

http://www.plantphenomics.org.au/projects/podd/ Online repository for storage and retrieval of raw and analyzed data from Australian Plant Phenomics Facility (APPF) phenotyping platforms. PODD is based on Fedora Commons repository software with data and metadata modeled using OWL/RDFS.

See also:

- PODD Project Site
  </small>

Plant Breeders API

Specifies a standard interface for plant phenotype/genotype databases to serve data for use in crop breeding applications. This is the API used by FieldBook, which allows users to turn spreadsheets into databases. Examples indicate that the responses will include values linked to the Crop Ontology, for example:

https://github.com/plantbreeding/API/blob/master/Specification/Traits/ListAllTraits.md

However, in general the BRAPI returned JSON data without linking context (i.e., not JSON-LD), so it is in essence it's own data structure.

Other notes:

- The Breeding Management System (BMS) group has implemented a few features to make it compatible with Field Book in its current state without the use of API.
- BMS and the Genomic & Open-source Breeding Informatics Initiative (GOBII) are both pushing for the API and plan on implementing it when it's complete.
- Read news about the BMS Breeding Management System Standalone Server and genomes2fields migrating to BMS

See also

- Plant Breeding API
  </small>

Plant Genomics and Phenomics Research Data Repository (PGP)

German repository for plant research data including image collections from plant phenotyping and microscopy, unfinished genomes, genotyping data, visualizations of morphological plant models, data from mass spectrometry as well as software and documents.

See also:

- Arend et al (2016). PGP repository: a plant phenomics and genomics data publication infrastructure. Database.
- PGP Repository
  </small>

USDA Plants

“The PLANTS Database provides standardized information about the vascular plants, mosses, liverworts, hornworts, and lichens of the U.S. and its territories. It includes names, plant symbols, checklists, distributional data, species abstracts, characteristics, images, crop information, automated tools, onward Web links, and references.”

See also

- USDA Plants Website
  </small>

USDA Quick Stats

Web based application supports querying the agricultural census and survey statistics. Also available via API.

See also

- USDA Quick Stats Website
  </small>

transPLANT

Infrastructure to support computational analysis of genomic data from crop and model plants. This includes the large-scale analysis of genotype-phenotype associations, a common set of reference plant genomic data, archiving genomic variation, and a search engine integrating reference bioinformatics databases and physical genetic materials. *See also*

- transPlant Website
  </small>

Sensor Data

Meteorological data

Proposed format for meteorological variables exported from Lemnatec platform

Multi-scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP) data formats

- One implementation of CF for ecosystem model driver (met, soil) and output (mass, energy dynamics)
  - Standardized Met driver data
  - Terrestrial Ecosystem Model output

Date-Time:

YYYY-MM-DD hh:mm:ssZ: based on ISO 8601 . Optional offset for local time; precision determined by data (e.g. could be YYYY-MM-DD and decimals specified by a period.

Agronomic and Phenotype Data Standards

Current Practice

In TERRA-REF v0 release, agronomic and phenotype data is stored and exchanged using the BETYdb API. Agronomic data is stored in the `sites`, `managements`, and `treatments` tables. Phenotype data is stored in the `traits`, `variables`, and `methods` tables. Data is ingested and accessed via the BETYdb API formats.

Standardization Efforts

In cooperation with participants from AgMIP, the Crop Ontology, and Agronomy Ontology groups, the TERRA-REF team is pursuing the development of a format to facilitate the exchange of data across systems based on the ICASA Vocabulary and AgMIP JSON Data Objects. An initial draft of this format is available for comment on Github.

In addition, we plan to enable the TERRA-REF databases to import and export data via the Plant Breeding API (BRAPI).

Genomic Data Standards

Overview

Genomic data have reached a high level of standardization in the scientific community. Today, all high-impact journals typically ask the author to deposit their genomic data in either or both of these databases before publication.

Below are the most widely accepted formats that are relevant to the data and analyses generated in TERRA-REF.

Raw reads + quality scores

Raw reads + quality scores are stored in FASTQ format. FASTQ files can be manipulated for QC with FASTX-Toolkit

Reference genome assembly

Reference genome assembly (for alignment of reads or BLAST) is in FASTA format. FASTA files generally need indexing and formatting that can be done by aligners, BLAST, or other applications that provide built-in commands for this purpose.

Sequence alignment

Sequence alignments are in BAM format – in addition to the nucleotide sequence, the BAM format contains fields to describe mapping and read quality. BAM files are binary files but can be visualized with IGV. If needed, BAM can be converted in SAM (text file) with SAMtools

BAM is the preferred format for sra database (sequence read archive).

SNP and genotype variants

SNP and genotype variants are in VCF format. VCF contains all information about read mapping and SNP and genotype calling quality. VCF files are typically manipulated with

vcftools

VCF format is also the format required by dbSNP, the largest public repository all SNPs.

Genomic coordinates

Genomic coordinates are given in a BED format – gives the start and end positions of a feature in the genome (for single nucleotides, start = end). BED files can be edited with bedtools.

See Also

- Genomics Data Pipeline
- Genomics Data Products

Sensor Data Standards

Current Practice

In the TERRA-REF release, sensor metadata is generally stored and exchanged using formats defined by LemnaTec. Sensor metadata is stored in `metadata.json` files for each dataset. This information is ingested into Clowder and available via the "Metadata" tab metadata.jsonld API endpoint.

Manufacturer information about devices and sensors are available via Clowder in the Devices and Sensors Information collection. This collection includes datasets representing each sensor or calibration target containing specifications\/datasheets, calibration certificates, and associated reference data.

Fixed metadata

Authoritative fixed sensor metadata is available for each of the sensor datasets. This has been extended to include factory calibrated spectral response and relative spectral response information. For more information, please see the sensor-metadata repository on Github.

Runtime metadata

Runtime metadata for each sensor run is stored in the `metadata.json` files in each sensor output directory.

Reference data

Additional reference data is available for some sensors:

- Factory calibration data for the LabSphere and SphereOptics calibration targets.
- Relative spectral response (RSR) information for sensors
- Calibration data for the environmental logger
- Dark/white reference data for the SWIR and VNIR sensors.

Standardization Efforts

The TERRA-REF team is currently investigating available standards for the representation of sensor information. Preliminary work has been done using OGC SensorML vocabularies in a custom JSON-LD context. For more information, please see the sensor-metadata repository on Github.

Data Standards Committee

The Standards Committee is responsible for defining and advising the development of data products and access protocols for the ARPA-E TERRA program. The committee consists of twelve core participants: one representative from each of the six funded projects and six independent experts. The committee will meet virtually each month and in person each year to discuss, develop, and revise data products, interfaces, and computing infrastructure.

Roles and responsibilities

TERRA Project Standards Committee representatives are expected to represent the interests of their TERRA team, their research community, and the institutions for which they work. External participants were chosen to represent specific areas of expertise and will provide feedback and guidance to help make the TERRA platform interoperable with existing and emerging sensing, informatics, and computing platforms.

Specific duties

- Participate in monthly to quarterly teleconferences with the committee.
- Provide expert advice.
- Provide feedback from other intersted parties.
- Participate in, or send delegate to, annual two-day workshops.

Annual Meetings

If we can efficiently agree on and adopt conventions, we will have more flexibility to use these workshops to train researchers, remove obstacles, and identify opportunities. This will be an opportunity for researchers to work with developers at NCSA and from the broader TERRA informatics and computing teams to identify what works, prioritize features, and move forward on research questions that require advanced computing.

People

Name	Institution	Email

Coordinators		
David Lee	ARPA-E	david.lee2_at_hq.doe.gov
David LeBauer	UIUC / NCSA	dlebauer_at_illinois.edu
TERRA Project Representatives		
Paul Bartlett	Near Earth Autonomy	paul_at_nearearthautonomy.com
Jeff White	USDA ALARC	Jeffrey.White_at_ars.usda.gov
Melba Crawford	Purdue	melbac_at_purdue.edu
Mike Gore	Cornell	mag87_at_cornell.edu
Matt Colgan	Blue River	matt.c_at_bluerivert.com
Christer Janssen	Pacific Northwest National Laboratory	georg.jansson_at_pnnl.gov
Barnabas Poczos	Carnegie Mellon	bapoczos_at_cs.cmu.edu
Alex Thomasson	Texas A&M University	thomasson_at_tamu.edu
External Advisors		
Cheryl Porter	ICASA / AgMIP / USDA	
Shawn Serbin	Brookhaven National Lab	sserbin_at_bnl.gov
Shelly Petroy	NEON	spetroy_at_neoninc.org
Christine Laney	NEON	claney_at_neoninc.org
Carolyn J. Lawrence-Dill	Iowa State	triffid_at_iastate.edu
Eric Lyons	University of Arizona / iPlant	ericlyons_at_email.arizona.edu

Data Product Levels

Data Product Levels

Data products are processed at various levels ranging from Level 0 to Level 4. Level 0 products are raw data at full instrument resolution. At higher levels, the data are converted into more useful parameters and formats. These are derived from NASA and NEON

Level	Description
0	Reconstructed, unprocessed, full resolution instrument data; artifacts and duplicates removed.
1a	Level 0 plus time-referenced and annotated with calibration coefficients and georeferencing parameters (level 0 is fully recoverable from level 1a data).
1b	Level 1a processed to sensor units (level 0 not recoverable)
2	Derived variables (e. g., NDVI, height, fluorescence) at the level 1 resolution.
3	Level 2 mapped to uniform grid, missing points gap filled; overlapping images combined
4	'phenotypes' derived variables associated with a particular plant or genotype rather than a spatial location

1 Earth Observing System Data Processing Levels, NASA
2 National Ecological Observatory Network Data Processing

Directory Structure

The data processing pipeline transmits data from origination sites to a controlled directory structure on the Nebula computer at NCSA where it is available for transfer via Globus. This directory structure is visible when accessing data via the Globus interface.

The data is generally structured as follows:

```
/sites
  /ua-mac
    /raw_data
      /sensor1
        /timestamp
          /dataset
      /sensor2
      ...
    /Level_1
      /extractor1_outputs
      /extractor2_outputs
      ...
  /danforth
    /raw_data
      /sensor3
      ...
    /Level_1
      /extractor3_outputs
```

...where raw outputs from sensors per site are stored in a `raw_data` subdirectory and corresponding outputs from different extractor algorithms are stored in `Level_1` (and `Level_2`, etc) subdirectories (see data product levels). When possible, sensor directories are divided into days and then into individual datasets.

Data Transfer

```
#Data Transfer
```

Maricopa Agricultural Center, Arizona

Environmental Sensors
Log of files transfered from Arizona to NCSA

Transferring ima

Data is sent to the gantry-cache server located inside the main UA-MAC building's telecom room via FTP over a private 10GbE interface. Path to each file being transferred is logged to /var/log/xferlog. Docker container running on the gantry-cache reads through this log file, tracking the last line it has read and scans the file regularly looking for more lines. File paths are scraped from the log and are bundled into groups of 500 to be transferred to the Spectrum Scale file systems that backs the ROGER cluster at NCSA via the Globus Python API. The log file is rolled daily and compressed to keep size in check. Sensor directories on the gantry-cache are white listed for being monitored to prevent accidental or junk data from being ingested into the Clowder pipeline.

A Docker container in the terra-clowder VM running in ROGER's Openstack environment gets pinged about incoming transfers and watches for when they complete, once completed the same files are queued to be ingested into Clowder.

Once files have been successfully received by the ROGER Globus endpoint, the files are then removed from the gantry-cache server by the Docker container running on the gantry-cache server. A clean up script walks the gantry-cache daily looking for files older than two days that have not been transferred and queues any if found.

Automated controlled-environment phenotyping, Missouri

Transferring images

Processes at Danforth monitor the database repository where images captured from the Scanalyzer are stored. After initial processing, files are transferred to NCSA servers for additional metadata extraction, indexing and storage.

At the start of the transfer process, metadata collected and derived during Danforth's initial processing will be pushed.

The current "beta" Python script can be viewed on GitHub. During transfer tests of data from Danforth's sorghum pilot experiment, 2,725 snapshots containing 10 images each were uploaded in 775 minutes (3.5 snapshots\/minute).

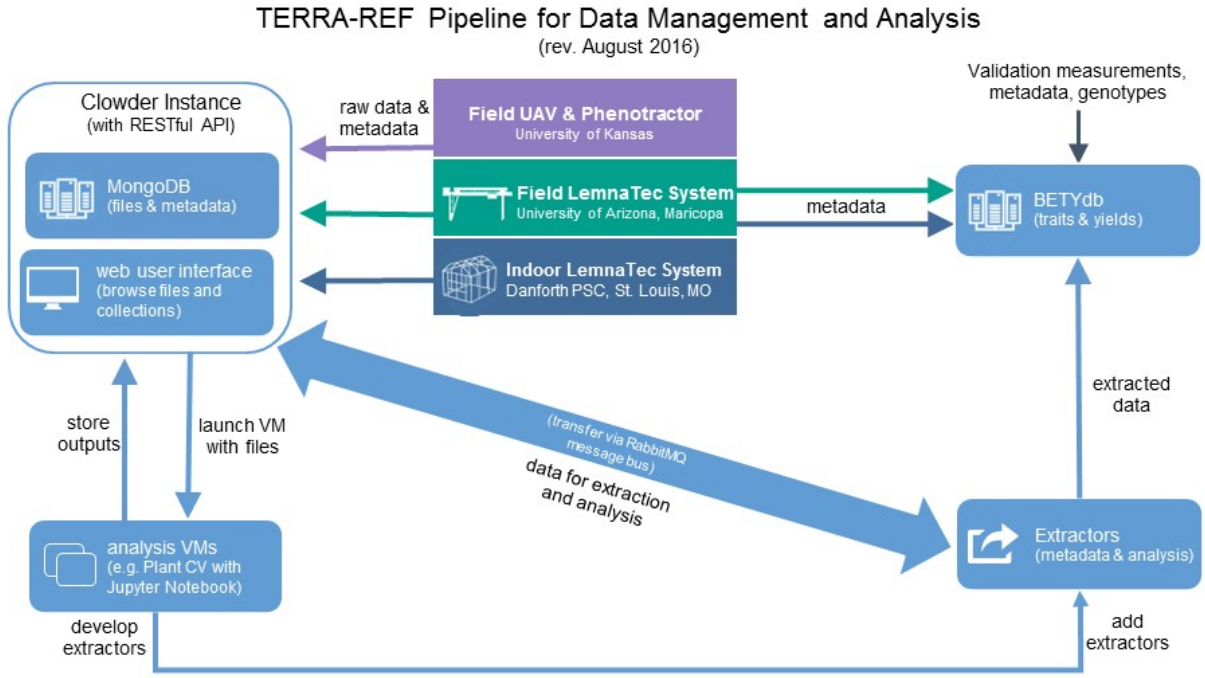
Transfer volumes

The Danforth Center transfers approximately X GB of data to NCSA per week.

Kansas State University

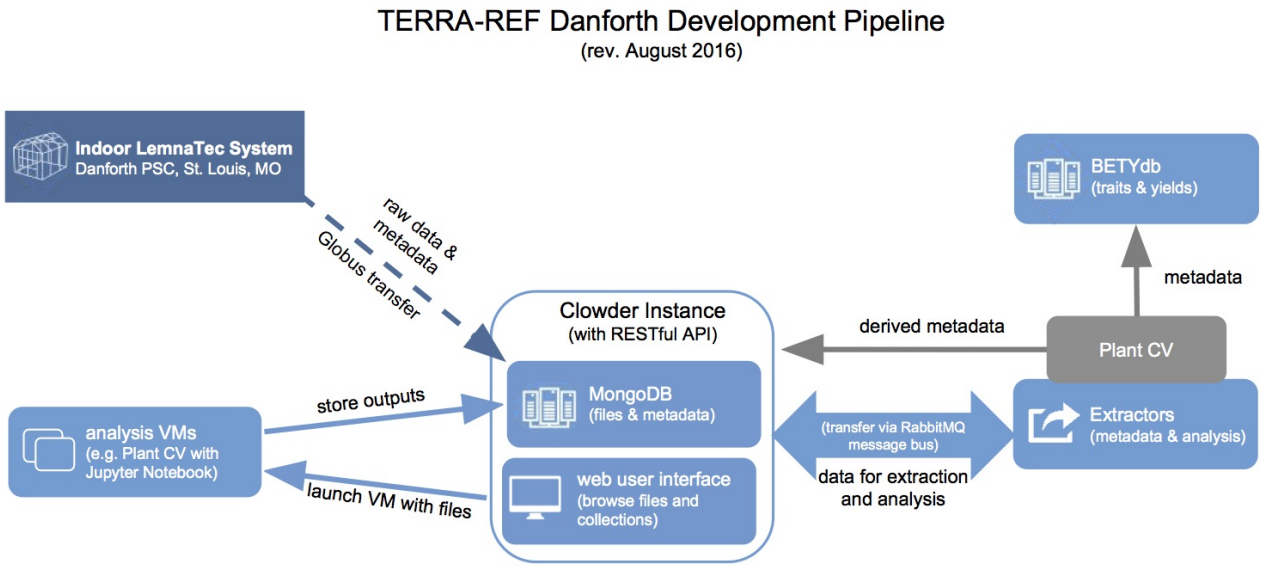
HudsonAlpha - Genomics

Data Processing Pipeline



Maricopa Agricultural Center, Arizona

Automated controlled-environment phenotyping, Missouri



At two points in the processing pipeline, metadata derived from collected data is inserted into BETYdb:

- At the start of the transfer process, metadata collected and derived during Danforth's initial processing will be pushed.
- After transfer to NCSA, extractors running in Clowder will derive further metadata that will be pushed. This is a subset of the metadata that will also be stored in Clowder's database. The complete metadata definitions are still being determined, but will likely include:
  - plant identifiers
  - experiment and experimenter
  - plant age, date, growth medium, and treatment
  - camera metadata

Kansas State University

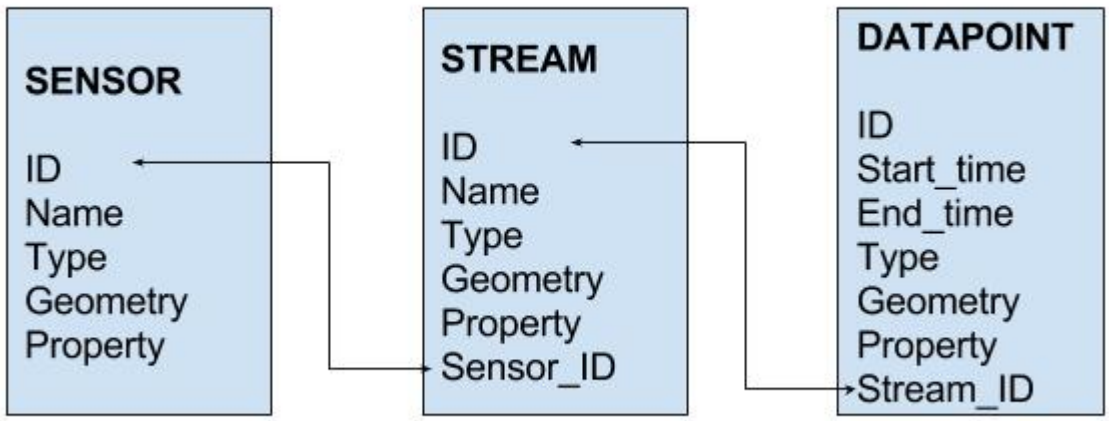
HudsonAlpha - Genomics

Time Series Data in Geostreams

Several extractors push data to the Clowder Geostreams database, which allows registration of data streams that accumulate datapoints over time. These streams can then be queried, visualized and downloaded to get time series of various measurements across plots and sensors. Learn more about data in this database in this tutorial

The TERRA-REF Geostreams database organizes data into three levels:

- Location (e.g. plot, or a stationary sensor)
    - Information stream (a particular instrument's data, or a subset of one instrument's data)
        - Datapoint (a single observation from the information stream at a particular point in time)



Generalized schema for the Geostreams database (part of clowder).

Sensor destinations

Here, the various streams that are used in the pipeline and their contents are listed.

- Location group

    - Stream name
        - **Datapoint property** [units / sample value]
        - ...
- Full Field (Environmental Logger)
    - Weather Observations
        - **sunDirection** [degrees / 358.4948271126]
        - **airPressure** [hPa / 1014.1764580218]
        - **brightness** [kilo Lux / 1.0607318339]
        - **relHumidity** [relHumPerCent / 19.3731498154]
        - **temperature** [DegCelsuis / 17.5243385113]
        - **windDirection** [degrees / 176.7864009522]
        - **precipitation** [mm/h / 0.0559327677]
        - **windVelocity** [m/s / 3.4772789697]
        - raw values shown here; check if extractor converts to SI units
    - Photosynthetically Active Radiation
        - **par** [umol/(m^2*s) / 0]
    - co2 Observations
        - **co2** [ppm / 493.4684409718]
    - Spectrometer Observations
        - **maxFixedIntensity** [16383]
        - **integration time in us** [5000]
        - **wavelength** [long array of decimals]
        - **spectrum** [long array of decimals]
- AZMET Maricopa Weather Station
    - Weather Observations
        - **wind_speed** [1.089077491]
        - **eastward_wind** [-0.365913231]
        - **northward_wind** [-0.9997966834]
        - **air_temperature** [Kelvin/301.1359779]
        - **relative_humidity** [60.41579336]
        - **preciptation_rate** [0]
        - **surface_downwelling_shortwave_flux_in_air** [43.60608856]
        - **surface_downwelling_photosynthetic_photon_flux_in_air** [152.1498155]
    - Irrigation Observations
        - **flow** [gallons / 7903]
- UIUC Energy Farm - CEN
- UIUC Energy Farm - NE

- UIUC Energy Farm - SE
  - Energy Farm Observations - CEN/NE/SE
    - **wind_speed**
    - **eastward_wind**
    - **northward_wind**
    - **air_temperature**
    - **relative_humidity**
    - **preciptation_rate**
    - **surface_downwelling_shortwave_flux_in_air**
    - **surface_downwelling_photosynthetic_photon_flux_in_air**
    - **air_pressure**
- PLOT_ID e.g. Range 51 Pass 2 (each plot gets a separate location group)
  - sensorName - Range 51 Pass 2 (each sensor gets a separate stream within the plot)
    - **fov** [polygon geometry]
    - **centroid** [point geometry]
  - canopycover - Range 51 Pass 2
    - **canopy_cover** [height/0.294124289126]

Data Backup

Raw data

Script uses the Spectrum Scale policy engine to find all files that were modified the day prior, and passes that list to a job in the batch system. The job bundles the files into a .tar file, then uses pigz to compress it in parallel across 18 threads. Since this script is run as a job in the batch system, with variables passed with the date, if the batch system is busy, the backups won't need to preclude each other. The .tgz files are then sent over to NCSA Nearline using Globus, then purged from file system.

BETYdb

Runs every night at 23:59. View the script.

This script creates a daily backup every day of the month. On Sundays creates a weekly backup, on the last day of the month it creates a monthly backup and at the last day of the year it will create a yearly backup. This script overwrite existing backups, for example every 1st of the month it will create a backup called bety-d-1 that contains the backup of the 1st of the month. See the script for the rest of the file names.

These backups are copied using crashplan to a central location and should allow recovery in case of a catastrophic failure.

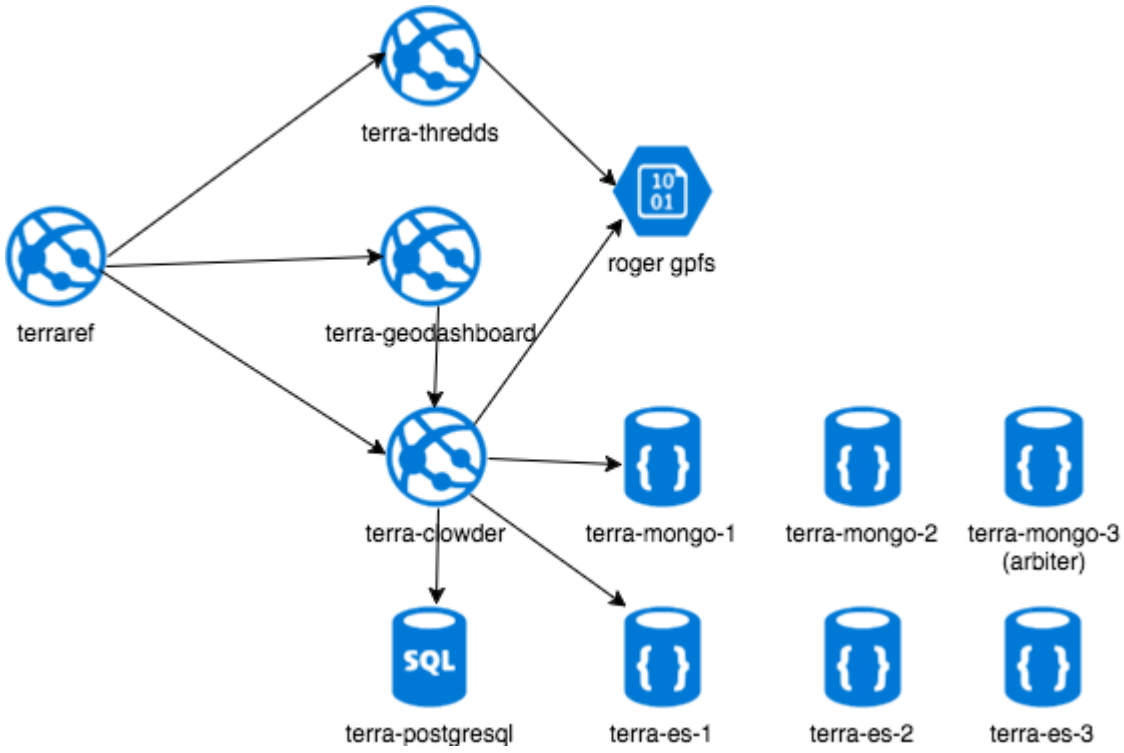
See Also

- Description of Blue Water's nearline storage system https://bluewaters.ncsa.illinois.edu/data
- Github issues:
  - https://github.com/terraref/computing-pipeline/issues/87
  - https://github.com/terraref/computing-pipeline/issues/384

Systems Configuration

The software that makes up the terraref system runs on different VM's. Some of the services leveraged by the systems runs in a replicated mode so that the overall system will not stop working if any of the underlying VM's goes down.

Following is the overview of the system as it is running now:



- terraref is the frontend for everything, runs nginx
- terra-clowder runs the data management system clowder, connected to terra-mongo-[123], terra-es-[123], terra-postgres and the NCSA storage condo (using NFS mount)
- terra-geodashboard runs the geodashboard software, connected to terra-clowder
- terra-thredds runs the thredds server (experimental), connected to roger filesystem (using NFS moutn)
- terra-es-[123] run elasticsearch 2.4 and for a cluster
- terra-mongo-[123] run mongo 3.6 in a replicated cluster, terra-mongo-3 is an arbiter and does not hold any data
- terra-postgres runs postgres 9.5

Developer Manual

TERRA members may submit data to Clowder, BETYdb, and CoGe.

- **Clowder** contains data related to the field scanner operations and sensor box, including bounding box of each image / dataset as well as location of the sensor, data types and processing level, scanner missions.
- **BETYdb** contains plot locations and other geolocations of interest (e.g. fields, rows, plants) that are associated with agronomic experimental design / meta-data (what was planted where, field boundaries, treatments, etc).
- **CoGe** contains genomic data.

They may also develop **extractors** - services that run silently alongside Clowder.

Submitting data to Clowder

Web Interface Data Uploads

1. Log in with your account
2. Click 'Datasets' > 'Create'
3. Provide a name and description
4. Click 'Select Files' to choose which files to add
5. Click 'Upload' to save selected files to dataset
6. Click 'View Dataset' to confirm. You can add more content with 'Add Files'.
7. Add metadata, terms of use, etc.

Some metadata may automatically be generated depending on the types of files uploaded. Metadata can be manually added to files or datasets at any time.

API Data Uploads

Clowder also includes a RESTful API that allows programmatic interactions such as creating new datasets and downloading files. For example, one can request a list of datasets using: GET _clowder home URL_/api/datasets. The current API schema for a Clowder instance can be accessed by selecting API from the ? Help menu in the upper-right corner of the application.

For typical workflows, the following steps are sufficient to push data into Clowder in an organized fashion:

1. Create a collection to hold relevant datasets (optional) `POST /api/collections` *provide a name; returns collection ID*
2. Create a dataset to hold relevant files and add it to the collection
   `POST /api/datasets/createempty` *provide a name; returns dataset ID*
   `POST /api/collections/<collection id>/datasets/<dataset id>`
3. Upload files and metadata to dataset
   `POST /api/datasets/uploadToDataset/<dataset id>` *provide file(s) and metadata*

An extensive API reference can be found *here*.

Uploading Data Using Globus

Some files, e.g. those transferred via Globus, will be moved to the server without triggering Clowder's normal upload paths. These must be transmitted in a certain way to ensure proper handling.

1. Log into *Globus* and click 'Transfer Files'.
2. Select your source endpoint, and Terraref as the destination. You need to contact NCSA to ensure you have the necessary credentials and folder space to utilize Globus - unrecognized Globus accounts will not be trusted.
3. Transfer your files. You will receive a Task ID when the transfer starts.
4. Send this Task ID and requisite information about the transfer to the TERRAREF Globus Monitor API as a JSON object:

```
{ "user": &lt;globus\_username&gt;
"globus\_id": &lt;Task ID&gt;
"contents": {
&lt;dataset1&gt;: {
&lt;filename1&gt;: {
"name": &lt;filename1&gt;,
"md": &lt;file\_metadata1&gt;
},
&lt;filename2&gt;: {"name": ..., "md": {...}},
&lt;filename3&gt;: {...},
...
},
&lt;dataset2&gt;: {
&lt;filename4&gt;: {...},
...
},
...
}
}
}
```

In addition to username and Task ID, you must also send a "contents" object containing each dataset that should be created in Clowder, and the files that belong to that dataset. This allows Clowder to verify it has handled every file in the Globus task.

5. The JSON object is sent to the API via an HTTP request:
   `POST 141.142.168.72:5454/tasks` For example, with cURL this would be done with:
   ```
   curl -X POST -u <globus_username>:<globus_password> -d <json_object>
   141.142.168.72:5454/tasks
   ```

In this way Clowder indexes a pointer to the file on disk rather than making a new copy of the file; thus the file will still be accessible via Globus, FTP, or other methods directed at the filesystem.

Submitting data to BETYdb

Submitting Data to BETYdb

BETYdb is a database used to centralize data from research done in all TERRA projects. (It is also the name of the Web interface to that database.) Uploading data to BETYdb will allow everyone on the team access to research done on the TERRA project.

Preliminary steps

Before submitting data to BETYdb, you must first have an account.

1. Go to the **BETYdb** homepage.
2. Click the "Register for BETYdb" button to create an account. If you plan to submit data, be sure to request "Creator" page access level when filling out the sign-up form.
3. Understand how the database is organized and what search options are avaible. Do this by exploring the data using the *Data* tab (see next section).

Exploring the data

The Data tab contains a menu for searching the database for different types of data. The Data tab is also the pathway to pages allowing you to add new data of your own. But if you have a sizable amount of trait or yield data you wish to submit, you will likely want to use the Bulk Upload wizard (see below).

As an example, try clicking the Data tab and selecting *Citations*, the first menu item. A page with a list of citations that have already been uploaded into the system appears.

Citations are listed by the *first author's last name*. For example a journal article written by Andrew Davis and Kerri Shaw would have the name "Davis" in the author slot.

Use the search box located in the top right corner of the page to search for citations by author, year, title, journal, volume, page, URL, or DOI. Note that the search string must exactly match a substring of the value of one of these items (though the matching is case-insensitive).

Each of the other collections listed in the Data menu may be searched similarly. For example, on the *Cultivars* page you can search cultivars in the system by searching for them by any of several facets pertaining to cultivars, including the name, ecotype, associated species, even the notes. Keep in mind that when switching to a new Data menu item (such as Cultivars), the resulting page will initially show all items of the type selected that are currently on file. (More precisely, since results are paginated, it will show the first twenty-five of those results.)

Preparing for bulk upload of data

The Bulk Upload wizard expects data in CSV format, with one row for each set of associated data items. ("Associated data items" usually means a set of measurements made on the same entity at the same time.) Each *trait* or *yield* data item must be associated with a citation, site, species, and treatment and *may* be associated with a specific cultivar of the associated species. Before you can upload data from a data file, this associated citation, site, species, cultivar, and treatment information must already be in place.

Moreover, if you are uploading *trait* data, your CSV data file must have one or more trait variable columns (and optionally, one or more *covariate* variable columns), and the names of these columns must match the names of existing variables. (See the discussion of variables below.)

Details on adding associated data

There is no bulk upload process for adding citations, site, species, cultivars, treatment, and variables to the database. They must be added one at a time using Web forms. Since most often a set of dozens or hundreds of traits is associated with a single citation, site, or species (etcetera), usually this is not an undue burden.

Details on checking that items of each particular type exist (and adding them if they don't) follow:

Citations: To check that the needed citations exist, go to the citations listing by clicking *Citations* in the Data menu. Search for your citation(s) to determine if all citations associated with your data already exist. If they don't, then create new citations as needed. Be sure to fill in all the required data; author, year, and title are *required*; if at all possible,

include the journal name, volume, page numbers, and DOI. (You *must* include the DOI if that is what your data files uses to identify citations.)

Sites: Go to the Data tab and click on *Sites* to verify that all sites in your data file are listed on the Sites page. If any of your sites are not already in the system, you will need to add them to the database. To do this, first search the citations list for the associated citation, select it (by clicking the checkmark in the row where it is listed) and then click the *New Site* button. A new site *must* have a name, but if possible, supply other information—the city, state, and country where the site is located, the latitude, longitude, and altitude of the site, and possibly climate and soil data.

It is possible that sites referenced by your data are already in the database but that they aren't yet associated with the citation associated with that data. To see the set of sites associated with a given citation, find the citation in the citations list and select it by clicking the checkmark in its row. This will take you to the *Listing Sites* page; all of the sites associated with the selected citation (if any) will be listed at the top. To associate another site with the selected citation, enter its name in the search box, find the row containing it, and click the "link" action in that row.

Treatments: The treatment specified for each of your data items must not only match the name of an existing treatment, it must also be associated with the citation for the data item. To see the list of treatments associated with a particular citation, select the citation as in the instructions for *Sites*. Then click the *Treatments* link on the *Listing Sites* page. The top section of this page lists all treatments associated with the selected citation.

Currently, there is no way to associate an arbitrary treatment with a citation via the Web interface. You will either have to make a new treatment with the desired name (after the desired citation has been selected), or you will have to (or have an administrator) modify the database directly.

Species: To check that the needed species entries exist, go to the the species listing by clicking *Species* in the Data menu. Search for each of the species required by your data. The species entry in the CSV file must match the scientific name (Latin name) of the species listed in the database. If necessary, add any species in your data that has not yet been added to the database. When adding a species, scientificname is the only *required* field, but the genus and species fields *should* be filled out as well.

Cultivars: If your data lists cultivars, you should check that these are in the database as well. Cultivar names are not necessarily unique, but they are unique within a given species. To check whether a cultivar matching the name and species listed in your CSV file has been added to the database, go to the cultivar listing by clicking *Cultivars* in the Data menu. Searching either by species name or cultivar name should quickly determine if the needed cultivar exists. If it needs to be added, click the *New Cultivar* button. Fill in the species search box with enough of the species name to narrow down the result list to a workable size, and then select the correct species from the result list immediately below the search box. Then type the name of the cultivar you wish to add in the *Name* field. The Ecotype and Notes sections are optional.

Variables: If you are submitting trait data, verify that the variables associated with each trait and each covariate match the names of variables in the system (for example, *canopy_height*, *hull_area*, or *solidity*). To do this, go to the Data tab and click on *Variables*. If any of your variables are not already in the system, you will need to add them.

For a variable to be recognized as a trait variable or covariate, it is not enough for it simply to be in the `variables` table; it must also be in the `trait_covariate_associations` table. To check which variables will be recogized as trait variables or covariates, click on the *Bulk Upload* tab. Then click the link *View List of Recognized Traits*. This will bring up a table that lists all names of variables recognized as traits and the names of all variables recognized as required or optional covariates for each trait. If you need to add to this table and do not have direct access to the underlying database to which you are submitting data, you will need to e-mail the site adminstrator to request additions. (See the "Contact Us" section in the footer of the **BETYdb** homepage.)

The Bulk Upload Wizard

Once you have entered all the necessary data to prepare for a bulk data upload, you can then begin the bulk upload process.

There are some key rules for bulk uploading:

1. **Templates** To help you get started, some data file templates are available. There are four different templates to choose from.

- yields_template_by_citation_author_year_title.csv
  Use this template if you are uploading yields and you wish to specify the citations by author, year, and title.
- yields_template_by_citation_doi.csv
  Use this template if you are uploading yields and you wish to specify the citations by DOI.
- traits_template_by_citation_author_year_title.csv
  Use this template if you are uploading traits and you wish to specify the citations by author, year, and title.
- traits_template_by_citation_doi.csv
  Use this template if you are uploading traits and you wish to specify the citations by DOI.

These "templates" consist of a single line of text showing a typical header row for a CSV file. In the traits templates, the headings of the form "[trait variable 1]" or "[covariate 1]" must be replaced with actual variable names corresponding to a trait variable or covariate, respectively.

These templates show all possible columns that may be included. In most cases, fewer columns will be needed and the unneeded column headings should be removed. The only programmatically *required* headings are "yield" (for uploads of yield data), or, for uploads of trait data, the name of at least one recognized trait variable. All other data required for an upload—the citation, site, species, treatment, access level, and date—may be specified interactively, provided that they have a uniform value for all of the trait or yield data in the file being uploaded. (Specification of a cultivar is not required, but it too may be specified interactively if it has a uniform value for all of the data in the file.)

2. **Matching** It is important that text values and trait or covariate column names in the data file match records in the database. This includes variable names, site names, species and cultivar names, etc. Note, however, that matching is somewhat lax: the matching is done case-insensitively, and extraneous spaces in values in the data file are ignored.

   Some special cases of note: In the case of `citation_title`, the supplied value need only match an initial substring of the title specified in the database as long as the combination of author, year, and the initial portion of the title uniquely identifies a citation stored in the database. (The value for `citation_title` may even be *empty* if the author and year together uniquely identify a citation!) And in the case of species names, the letter 'x' may be used to match the times symbol '×' used in names of hybrid species.

3. **Column order** The order of columns in the data file is immaterial; in making the template files, an arbitrary order was chosen. But because the data in the data file is displayed for review during the bulk upload process, it may be that some orderings are easier to work with than others.
4. **Quotation rules** Since commas are used to delineate columns in CSV files, any data value containing a comma must be surrounded by double quotes. (Single quotes are interpreted as part of the value!) If the value itself contains a double-quote, this double-quote must be doubled ("") in addition to surrounding the value with double quotes.
5. **Character encoding** Non-ASCII characters must use UTF-8 encoding.
6. **Blank lines** There can be no blank lines in the file, either between data rows or at the end of the file.

Troubleshooting data files

Immediately after uploading a data file (or after specifying the citation if this is done interactively), the Bulk Upload Wizard tries to validate the uploaded file and displays the results of this validation.

The types of errors one may encounter at this stage fall into roughly three categories:

1. Parsing errors
   These are errors at the stage of parsing the CSV file, before the header or data values are even checked. An error at this stage returns one to the *file-upload* page.
2. Header errors
   These are errors caused by having an incongruous set of headings in the header row. Here are some examples:
   1. There is `citation_author` column heading without a corresponding `citation_year` and `citation_title` heading. It is an error to use one of these headings without the other two.
   2. There is both a `citation_doi` heading and a `citation_author`, `citation_year`, or `citation_title` heading. If `citation_doi` is used, none of the other citation-related headings is allowed.
   3. There is an `SE` heading without an `n` heading or vice versa.
   4. There is neither a `yield` heading nor a heading corresponding to a recognized trait variable.
   5. There is both a `yield` heading and a heading corresponding to a recognized trait variable. A data file can be used to insert data into the traits table or the yields

table but not both at once.

6. There is a `cultivar` heading but no `species` heading.

If any of these errors occur, validation of data values will not proceed.

There may be other errors associated with the header row that aren't treated as errors as such. For example, if you intend to supply two trait variables per row but misspell one of them, the data in the column headed by the misspelled variable name will simply be ignored. That column will be grayed-out, but the file may still be used to insert data corresponding to the "good" variable (provided there are no other errors). In other words, if you ignore the "ignored column" warning and the gray highlighting, you may end up uploading only a portion of the data you intended to upload.

3. Value errors

   If there are no file-parsing errors or header errors, the Bulk Upload wizard will proceed to validate data values. Valid values will be highlighted in green. Ignored columns will be highlighted in gray. (This will warn you, for example, if you have misspelled the name of a trait variable.) Other colors signify various sorts of errors. A summary of errors is shown at the top of the page with links to rows in which the various errors occur.

   1. Matching value errors

      Each row of the CSV file must be associated with a unique citation, site, species, and treatment and *may* be associated with a unique cultivar. These associations may either be specified in the CSV file or, if a particular association is constant for all rows of the file, it may be specified interactively. If they *are* specified in the file, problems that may arise include:

      - The combination of values for `citation_author`, `citation_year`, and `citation_title` do not uniquely identify a citation in the database. (This may be because there are no matches or too many (i.e., more than one) matches. (There should never be multiple database rows having the same combination of author, year, and title, but this is not currently enforced.)
      - The value for `citation_doi` does not uniquely match a citation in the database. (Again, citation DOIs *should* be unique, but the database schema doesn't enforce this.)
      - The value for `site` does not uniquely match the sitename of a site in the database. (`site.sitename` *should* be unique, but this again is not enforced.)
      - The site specified in a given row is not consistent with the citation specified in that row. (If you visit the "Show" page for the site, you should see the citation listed at the top of the page right under *Viewing Site*.)

- The value for `species` does not match the value of `scientificname` for a unique row of the species table. (`species.scientificname` should be unique, but the database scheme doesn't currently enforce this.)
- The value for `treatment` does not match the value of the name of any treatment row in the database.
- The value for `treatment` in a particular row matches one or more treatments in the database, but none are associated with the citation specified by that row.
- The value for `treatment` in a particular row matches more that one treatment in the database that is associated with the citation specified by that row. (This error is rare. Names of treatments associated with a particular citation should be unique, but this is not yet enforced.)
- The value for `cultivar` specified in a particular row is not consistent with the species specified in that row.

2. Other value errors, not having to do with associated attributes of the data, are as follows:
   - A value for a trait is out of range. An obvious example would be giving a negative number as the value for annual yield. If a variable value is flagged as being out of range, double check the data. If you determine that the value is indeed correct, you should request to have the range in the database adjusted for that variable.
   - A value for the measurement date is not in the correct format or is out of range.
   - A value for the access level is not 1, 2, 3, or 4.
   - A value of the wrong type is given. Examples would be giving a text value for `yield` or a floating point number for `n` .

After successful validation

Global options and values

If there are no errors in the data file, the bulk upload will proceed to a page allowing you to choose rounding options for your data values. You may choose to keep 1, 2, 3, or 4 significant digits, 3 being the default. If your data includes a standard error (`SE`) column, you may separately specify the amount of rounding for the standard error. Here the default is 2 significant digits.

If you did not specify all associated-data values and or did not specify an *access level* in the data file itself, this page will also allow you to specify a uniform global value for any association not specified in the file; and it will allow you to specify a uniform access level if your data file did not have an `access_level` column.

Verification page

Once you have specified global options and values, you will be taken to a verification page that will summarize the global options you have selected and the associations you specified for your data. The latter will be presented in more detail than any specification in your data file or on the *Upload Options and Global Values* page. For example, when summarizing the sites associated with your data, not only are the site names listed, but the city, state, country, latitude, longitude, soil type, and soil notes are also displayed. This will help ensure that the citations, sites, species, etc. that you specified are really the ones that you intended.

Once you have verified the data, clicking the *Insert Data* button will complete the upload. The insertions are done in an SQL transaction: if any insertion fails, the entire transaction is rolled back.

Submitting Data to CoGe

CoGe supports the genomics pipeline required for the TERRA program for Sorghum sequence alignment and analysis. It has a web interface and REST API. CoGe is developed by Eric Lyons and hosted at the University of Arizona, where it is made available for researchers to use. CoGe can be hosted on any server, VM, or Docker container.

Submitting Sequences to the CoGe Pipeline

- Upload files to Cyverse data store. The TERRARef project has a 2TB allocation
- Use icommands to transfer to data store

CyVerse data store

- project directory: `/iplant/home/shared/terraref`
  - Raw data goes in subdirectory `raw_data/`, which is only writable for those sending raw reads.
- (CoGe output) can go into `output/`

Uploading data to data store using icommands

icommands documentation

Transferring data from Roger to iplant data store

```
# install icommands
cd $HOME
mkdir bin
cd bin
wget http://www.iplantcollaborative.org/sites/default/files/irods/icommand
```

```
tar -xvf icommands.x86_64.tar.bz2
# add icommands directory to $PATH
export PATH=$HOME/bin/icommands:$PATH
# initialize
iinit
# host name: data.iplantcollaborative.org
# port number:1247
# user name:(your Cyverse Login)
# Enter your irods zone:iplant
# iRODS password:*******
icd /iplant/home/shared/terraref/raw_data/hudson-alpha/
## transfer test data to iplant data store
touch checkpoint-file
iput -P -b -r -T --retries 3 -X checkpoint-file test_data/
```

Developing Clowder Extractors

Developing Clowder Extractors

Developing the Computing Pipeline with Clowder Extractors

The TERRA REF computing pipeline and data management is managed by Clowder. The pipeline consists of 'extractors' that take a file or other piece of information and generate new files or information. In this way, each extractor is a step in the pipeline.

An extractor 'wraps' an algorithm in code that watches for files that it can convert into new data products and phenotypes. These extractors wait silently alongside the Clowder interface and databases. Extractors can be configured to wait for specific file types and automatically execute operations on those files to process them and extract metadata.

If you want to add an algorithm to the TERRAREF pipeline, or use the Clowder software to manage your own pipeline, extractors provide a way of automating and scaling the algorithms that you have. The NCSA Extractor Development wiki provides instructions, including:

1. Setting up a pipeline development environment on your own computer.
2. Using the web development interface) (currently in beta testing)
3. Using the Clowder API
4. Using the pyClowder pyClowder pyClowder to add an analytical or technical component to the pipeline.

What does it take to contribute an extractor?

Overview

The purpose of this document is to define the requirements for contributing and maintaining algorithms to the TERRA REF pipeline.

How does an extractor developer get from drafting to deploying an extractor?

The stereo-rgb extractor is a good example of a completed extractor:

- https://github.com/terraref/extractors-stereo-rgb

ISDA has an overview of some common Python conventions for reference:

- https://docs.google.com/document/d/1n8iQHdb32u0EOkiNQSRK51XlAjGDsxg75W2SUAAMouM/edit#heading=h.wd4g4fd6q72u

Roles

- Science Developer (e.g. Zongyang, Sean, Patrick)
  - Writes, tests, documents science code
  - Works with pipeline developer to integrate and deploy
  - Works with end users of data to assess quality
- Pipeline Developer / Operator (e.g. Max, Todd)
  - Develops workflow code
  - Maintains real-time processing
  - Coordinates annual re-processing
- End User
  - Scientist who will be using the output data
  - Defines specifications
  - Identifies data that can be used for calibration and validation
  - Reviews output during development and continuous operation

The Extractor Lifecycle

Lets define three stages of extractor development. This is iterative, and there should be open communication among the Science Developer, Pipeline Developer, and End User throughout the process.

1. Define the extractor
   - Create an issue in Github to track development (information can later be added to README file)

    - Inputs (with examples)
    - Outputs
        - Add (or use) a citation, variable, and method in BETYdb
    - Data for ground truthing, testing, validation
2. Draft the extractor
    - Create a working 'feature' branch on GitHub
    - This should be updated regularly; this helps collaborators keep up to date
    - Use docstring for inline documentation https://www.python.org/dev/peps/pep-0257
3. Request feedback on initial draft and sample output
    - From Pipeline Developer
    - From End User
    - Revise based on feedback
4. Beta Release
    - Create a Pull Request when extractor is ready to deploy. The PR should be reviewed by both the Pipeline Operator and End User, who will either request changes or approve the PR.
    - A complete extractor is defined below
5. Deployment
    - Extractor deployed
        - First on live data stream. Data should indicate beta status of extractor
        - Then for reprocessing
    - Extractor added to the list in gitbook
    - Example of how to access actual output generated by extractor (e.g. BETYdb API call)
    - Versioned and pushed to PyPi if science package was extended
6. Operation
    - Output of extractor is vetted both by domain expert and code provider
    - Improvement

When is an extractor ready to be deployed?

All of the following are required for an extractor to be considered 'complete':

1. Expected test input
    - Expected test input may either be placed in in repository if 1MB, place the test input to Globus or under the tests/ directory in the Workbench..

    - This should include both real and simulated data representing a range of successful and failure conditions
2. Expected test output
3. Implementation
4. Example of output
5. Output is vetted by domain expert
6. Wrapped as extractor
7. Inline documentation with docstrings https://www.python.org/dev/peps/pep-0257
8. Documentation in README
    - Authors
        - One should be identified as maintainer / point of contact
    - Overview
        - Description
        - Inputs
        - outputs
    - Implementation (algorithm details)
        - Libraries used
        - References
        - Rationale (e.g. why method x over y)
    - QA/QC
        - Automated checks done in real time
        - Failure conditions
        - Known issues
    - Further Reading and Citations
        - Related Github issues
        - References
9. Documentation in extractor_info.json with documentation (maybe use @FILE to read file into json document)

TERRA-REF Extractor Resources

terrautils

To make working with the TERRA-REF pipeline as easy as possible, the terrautils Python library was written. By importing this library in an extractor script, developers can ensure that code duplication is minimized and standard practices are used for common tasks such

as GeoTIFF creation and georeferencing. It also provides modules for managing metadata, downloading and uploading, and BETYdb/geostreams API wrapping.

Modules include:

- betydb BETYdb API wrapper
- extractors General extractor tools e.g. for creating metadata JSON objects and generating folder hierarchies
- formats Standard methods for creating output files e.g. images from numpy arrays
- gdal GDAL general image tools
- geostreams Geostreams API wrapper
- influx InfluxDB logging API wrapper
- lemnatec LemnaTec-specific data management methods
- metadata Getting and cleaning metadata
- products Get file lists
- sensors Standard sensor information resources
- spatial Geospatial metadata management

Science packages

To keep code and algorithms broadly applicable, TERRA-REF is developing a series of science-driven packages to collect methods and algorithms that are generic to an input and output from the pipeline. That is, these packages should not refer to Clowder or extraction pipelines, but instead can be used in applications to manipulate data products. They are organized by sensor.

These packages will also include test suites to verify that any changes are consistent with previous outputs. The test directories can also act as examples on how to instantiate and use the science packages in actual code.

- stereo_rgb stereo RGB camera (stereoTop in raw*data, rgb* prefix elsewhere)
- flir_ir FLIR infrared camera (flirIrCamera in raw*Data, ir* prefix elsewhere)
- scanner_3d laser 3D scanner (scanner3DTop in raw*Data, laser3d* elsewhere)

Extractor repositories

Extractors can be considered wrapper scripts that call methods in the science packages to do work, but include the necessary components to communicate with TERRA's RabbitMQ

message bus to process incoming data as it arrives and upload outputs to Clowder. There should be no science-oriented code in the extractor repos - this code should be implemented in science packages instead so it is easier for future developers to leverage.

Each repository includes extractors in the workflow chain corresponding to the named sensor.

- extractors-stereo-rgb
- extractors-3dscanner
- extractors-multispectral
- extractors-metadata
- extractors-hyperspectral
- extractors-environmental
- extractors-lemnatec-indoor

Contact:

- Extractor development and deployment: Max Burnette
- Development environments: Craig Willis
- On our Slack Channel
- On GitHub

Code of Conduct

As contributors and maintainers of this project, we pledge to respect all people who contribute through reporting issues, posting feature requests, updating documentation, submitting pull requests or patches, and other activities.

We are committed to making participation in this project a harassment-free experience for everyone, regardless of level of experience, gender, gender identity and expression, sexual orientation, disability, personal appearance, body size, race, ethnicity, age, or religion.

Examples of unacceptable behavior by participants include the use of sexual language or imagery, derogatory comments or personal attacks, trolling, public or private harassment, insults, or other unprofessional conduct.

Project maintainers have the right and responsibility to remove, edit, or reject comments, commits, code, wiki edits, issues, and other contributions that are not aligned to this Code of Conduct. Project maintainers who do not follow the Code of Conduct may be removed from the project team.

This code of conduct applies both within project spaces and in public spaces when an individual is representing the project or its community.

Instances of abusive, harassing, or otherwise unacceptable behavior may be reported by opening an issue or contacting one or more of the project maintainers.

This Code of Conduct is adapted from the Contributor Covenant, version 1.1.0, available from http://contributor-covenant.org/version/1/1/0/

Appendix

Glossary

Accession - plant materials collected from a particular area.

Active reflectance - measurement of light originating from a sensor that reflects off of an object and back to the sensor

Algorithm - a process or set of rules to be followed in calculations or other problem-solving operations

Alignment, sequence - a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences

API (application programming interface) - a set of routine definitions, protocols, and tools for building software and applications.

BAM (Binary Alignment/Map) format - binary format for storing sequence data.

BED (Browser Extensible Data) format - format consisting of one line per feature, each containing 3-12 columns of data, plus optional track definition lines.

BETYdb (Biofuel Ecophysiological Traits and Yields database) - a web-based database of plant trait and yield data that supports research, forecasting, and decision making associated with the development and production of cellulosic biofuel crops

BRDF (Bidirectional Reflectance Distribution Function) - a function of four real variables that defines how light is reflected at an opaque surface.

Breeding Management System (BMS) - an information management system developed by the Integrated Breeding Platform to help breeders manage the breeding process, from program planning to decision-making.

Brown Dog - a research project to develop a method for easily accessing historic research data stored in order to maintain the long-term viability of large bodies of scientific research.

BWA - a software package for mapping low-divergent sequences against a large reference genome.

Clowder - a scalable data repository for sharing, organizing and analyzing data

Collections - one or more datasets.

Cultivar - plants selected for desirable characteristics that can be maintained by propagation.

Data product level - relative amount that data products are processed. Level 0 products are raw data at full instrument resolution. At higher levels, the data are converted into more useful parameters and formats.

Data standards - the rules by which data are described and recorded.

Datasets - one or more files with associated metadata collected by one sensor at one time point.

Downwelling spectral irradiance - The component of radiation directed toward the earth's surface per unit frequency or wavelength

Exposure - the amount of light per unit area reaching an electronic image sensor

FASTQ format - a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores.

FASTX-toolkit - a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

Gantry - a rail-bound crane systems that transport a measurement platform (like the Scanalyzer) over a field

GAPIT (Genome Association and Prediction Integrated Tool) – an R package that performs Genome Wide Association Study (GWAS) and genome prediction (or selection).

GATK (Genome Analysis Toolkit) - a software package for analysis of high-throughput sequencing data

Gbrowse - a combination of database and interactive web pages for manipulating and displaying annotations on genomes.

Generic Model Organism Database (GMOD) - a collection of open source software tools for managing, visualizing, storing, and disseminating genetic and genomic data.

Genome annotation - the process of attaching biological information to sequences.

Genomic coordinates - The beginning and ending positions of an annotation along a sequence

Genotype calling - inferring the genotype carried by an individual at each site

GeoDjango - geographic Web framework for building GIS Web applications

Germplasm - the sum total of genetic resources of an organism.

GFF (General Feature Format) - format consisting of one line per feature, each containing 9 columns of data, plus optional track definition lines

GIS (geographic information system) - a system designed to capture, store, manipulate, analyze, manage, and present all types of spatial or geographical data.

Globus - a connected set of data transfer and sharing services for research data management.

Hierarchical Data Format (HDF) - a set of file formats (HDF4, HDF5) designed to store and organize large amounts of data.

Hyperspectral data - information from across the electromagnetic spectrum.

IGV (Integrative Genomics Viewer) - a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.

Integrated Breeding Platform (IBP) - platform providing integrated, high-performing breeding informatics and management system

Jbrowse - an embeddable genome browser

Json - open-standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs.

Jupyter Notebook - a web application for creating and sharing documents that contain live code, equations, visualizations and explanatory text.

Lemnatec - supplier of software and automated research platforms for plant phenotyping.

Metadata - data that provides information about other data

MLMM (multi-locus mixed-model) - analysis for genome-wide association studies (GWAS) that uses a forward and backward stepwise approach to select markers as fixed effect covariates in the model.

NetCDF - a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data.

OpenAlea - a distributed collaborative effort to develop Python libraries and tools that address the needs of current and future works in Plant Architecture modeling.

OpenCV (Open Source Computer Vision Library) - an open source computer vision and machine learning software library.

PAR (Photosynthetically Active Radiation) - the amount of light available for photosynthesis, which is light in the 400 to 700 nanometer wavelength range.

Phenotype - the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment.

Phytozome - a project that facilitates comparative genomic studies amongst green plants.

PlantCV - an imaging processing package specific for plants that is built upon open-source software

PostGIS - an open source software program that adds support for geographic objects to the PostgreSQL object-relational database.

Python - a programming language

QA (quality assurance) - a planned system of review procedures conducted outside the actual data compilation.

QC (quality control) - a system of checks to assess and maintain the quality of the data.

Quality scores - measure of the probability that a nucleotide base is correctly identified from DNA sequencing

R/qtl - an extensible, interactive environment for mapping quantitative trait loci (QTL) in experimental crosses.

Raw data - unprocessed data collected from an experiment

Reads - sequence of nucleotides of a segment of DNA

Reference data - data that defines the set of permissible values to be used by other data fields.

RESTful API - an application program interface (API) that uses HTTP requests to get, put, post, and delete data.

ROGER - a cluster housed at NCSA that has 13.3 TB of system memory available for computation

Rstudio - a set of integrated tools for use with R, a software environment for statistical computing and graphics.

SAMtools (Sequence Alignment/Map) – a generic format for storing large nucleotide sequence alignments.

Scanalyzer - instrumentation created by Lemnatec with robotic sensor arm with multiple overhead cameras and sensors

Sequencing - the process of determining the precise order of nucleotides within a DNA molecule.

SNP (single nucleotide polymorphism) - a variation in a single nucleotide that occurs at a specific position in the genome

Spaces - contain collections and datasets. TERRA-REF uses one space for each of the phenotyping platforms.

Spectral exposure - the radiant energy received by a surface, per unit time, per unit frequency

Spectral flux - the radiant energy emitted, reflected, transmitted or received, per unit time, per unit frequency

Spectral response function (SRF) - the quantum efficiency of a sensor at specific wavelengths over the range of a spectral band

SQL (Structured Query Language) is a special-purpose programming language designed for managing data held in a relational database management system

SRA (Sequence Read Archive) - a bioinformatics database that provides a public repository for DNA sequencing data

Standards committee - TERRA project representatives and external advisors who work to create clear definitions of data formats, semantics, and interfaces, file formats, and representations of space, time, and genetic identity based on existing standards, commonly used file formats, and user needs to make it easier to analyze and exchange data and results.

Swagger - a set of rules for a format describing REST API. The format can be used to share documentation among product managers, testers and developers, but can also be used by various tools to automate API-related processes.

TASSEL-GBS - software for investigating the relationship between phenotypes and genotypes

TERRA (Transportation Energy Resources from Renewable Agriculture) - a program funded by ARPA-E program that facilitates the improvement of advanced biofuel crops, by developing and integrating cutting-edge remote sensing platforms, complex data analytics tools, and high-throughput plant breeding technologies.

TERRA-REF (Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform) - a research project focused on developing an integrated phenotyping

system for energy sorghum that leverages genetics and breeding, automation, remote plant sensing, genomics, and computational analytics.

Thredds: Geospatial Data server - a web server that provides metadata and data access for scientific datasets, using a variety of remote data access protocols

Trait - the morphological, anatomical, physiological, biochemical and phenological characteristics of plants and their organs

Variants - a nucleotide difference in a genotype compared to a reference genotype

VCF - a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome.

Vcftools - a program package designed for working with VCF files

White reference, reflectance of - light reflecting off of a white reference object that is used for the calibration of hyperspectral images

Accessing BETYdb with GIS Software

Here we describe how to access BETYdb directly from GIS software such as ESRI ArcMap and QGIS in order to query BETYdb data. The following instructions assume you have followed the instructions for setting up a copy of the TERRA REF database on your own machine described in the BETYdb section of How to Access Data.

The configurations used by QGIS and ArcMAP should be consistent with other software that uses databases.

Add BETYdb Layer or Table to ArcMap

BETYdb is configured with PostGIS geometry support. This allows ArcGIS Desktop clients to access geometry layers stored within BETYdb.

1. Click on the ArcCatalog icon (on right edge of ArcMap window) to open the ArcCatalog Tree
2. In the tree, click on 'Database Connections' and then "Add Database Connnections". A Database Connection dialog window will open.
3. Within the dialog box:

```
Database Platform: PostgreSQL
Instance: localhost
Authentication Type: Database authentication
User name: bety
Password: bety
Database: select bety (if everything else is correct)
```

4. Click OK
5. The connection will be saved as "Connection to localhost.sde", right click and rename to it to "TERRA REF BETYdb trait database" to allow easy reuse.
6. Click on the Add Layer icon (black cross over yellow diamand) button to open the Add Data dialog window.
7. Under 'Look in' on the second line choose 'Database Connections'.
8. Select the "TERRA REF BETYdb trait database" that created above

9. Select the bety.public.sites table and click 'Add'.
    - This 'sites' table is the only table in the database with a geospatial 'geometry' data type.
    - Any of the other tables can also be added, as described below.
10. The New Query Layer dialog will be displayed asking for the Unique Identifier Field for the layer. For the bety.public.sites table, the unique identifier is the "sitename" field.
11. Click Finish.

> *Warning*: ArcMap does not support the big integer format used by BETYdb as primary keys and those fields will not be visible or available for selection. In most cases you should be able to use other fields as unique identifiers.*

Modifying the Query Layer

BETYdb contains one geometry table called betydb.public.sites containing the boundaries for each plot. Because the plot boundaries can change each season, and even within season, different
plot definitions may be used (e.g. to subset plots or exclude boundary rows), there is significant overlap that can cause confusion
when displayed.
In general, you will want to use the query layer to limit plots to a single season and a single definition.

1. Right click the bety.public.sites layer and choose properties.
2. Choose the Definition Query tab
3. Add the line `sitename LIKE 'MAC Field Scanner Season 1%'` or `sitename LIKE 'MAC Field Scanner Season 2%'` to limit the layer to Season 1 or Season 2 respectively.
4. Click 'OK'

For more advanced selection of sites by experiment or season, you can join the `experiments` and `experiments_sites` tables. This is beyond the scope of the present tutorial.

Joining Additional BETYdb Tables

Additional tables can be added and joined to the sites table. Tables can be added just like any other layer. In this case, we'll add bety.public.traits_and_yields_view and join it to the bety.public.sites layer.

1. To create a join with other tables, start by adding the desired table.
2. Follow instructions above to add the bety.public.traits_and_yields_view
3. On this table the unique identifier is a group of columns, so select sitename, cultivar, scientificname, trait, date, entity, and method as the unique identifiers.
4. Right click on the bety.public.sites layer.
5. Under 'Joins and Relates' select 'Join'.
6. Choose sitename (from bety.public.sites) in part 1
7. Choose bety.public.traits_and_yields_view in part 2
8. Choose sitename in part 3
9. Click OK

Creating a Thematic View

The final section describes how to create a thematic view of the bety.public.sites layer based on the mean attribute where the trait is NDVI from the bety.public.traits_and_yields_view. Remove any previous joins from bety.public.sites (right click bety.public.sites –> joins and relates –> remove join) prior to performing this procedure because we will be selecting the NDVI data by creating a query layer from bety.public.traits_and_yields_view prior to the join.

1. Right click bety.public_traits_and_yields_view table and select properties
2. Click on the Definition Query tab
3. Add the line "trait = 'NDVI'" to the Definition Query box
4. Click OK
5. Follow the steps defined in Joining Additional BETYdb Tables
6. Right click on the bety.sites layer and choose properties
7. Choose the Symbology tab
8. Under the Show section, choose Quantities –> Graduated Colors
9. Under the Fields Value selection choose mean
10. Click OK

Connecting to Other GIS Software

Below connection instructions assume an SSH tunnel exists.

ArcGIS Pro

This assumes you have followed instructions for ArcMAP to create a database connection file.

- Open ArcCatalog
  - Under database connections, you will find the connection made above, called 'TERRA REF BETYdb.sde'
  - right click this and select 'properties'
  - copy the file path (it should look like
    ```
    C:\Users\<USER NAME>\AppData\Roaming\ESRI\Desktop10.4\ArcCatalog\TERRA REF BETYdb.sde
    ```
- Open ArcGIS Pro
  - Under the Insert tab, select connections –> 'add database'
  - paste the path to 'TERRA REF BETYdb.sde' in the directory navigation bar
  - select 'TERRA REF BETYdb.sde'

QGIS

- Open QGIS
- In left 'browser panel', right-click the PostGIS icon
- select 'New Connection'
- Enter connection properties
  - Name: TERRA REF BETYdb trait database
  - Service: blank
  - Host: localhost
  - Port: 5432
  - Database: bety
  - SSL mode: disable
  - Username: bety
  - Password: bety
  - Options: select 'Also list tables with no geometry'

How to export plots from PostGIS as a Shapefile

This does not require GIS software other than the PostGIS traits database. While connecting directly to the database within GIS software is handy, it is also straightforward to export Shapefiles.

After you have connected via ssh to the PostGIS server, the `pgsql2shp` function is available and can be used to dump out all of the plot and site definitions (names and geometries) thus:

```
pgsql2shp -f terra_plots.shp -h localhost -u bety -P bety bety \
          "SELECT sitename, geometry FROM sites"
```

References

Morris, Geoffrey P., Davina H. Rhodes, Zachary Brenton, Punna Ramu, Vinayan Madhumal Thayil, Santosh Deshpande, C. Thomas Hash et al. "Dissecting genome-wide association signals for loss-of-function phenotypes in sorghum flavonoid pigmentation traits." G3: Genes, Genomes, Genetics 3, no. 11 (2013): 2085-2094.
https://doi.org/10.1534/g3.113.008417

Reflectance Indices

Henrich, V., Krauss, G., Götze, C., Sandow, C. (2012): **IDB - www.indexdatabase.de, Entwicklung einer Datenbank für Fernerkundungsindizes.** AK Fernerkundung, Bochum, 4.-5. 10. 2012. (PDF)

Henrich, V., Jung, A., Götze, C., Sandow, C., Thürkow, D., Gläßer, C. (2009): **Development of an online indices database: Motivation, concept and implementation.** 6th EARSeL Imaging Spectroscopy SIG Workshop Innovative Tool for Scientific and Commercial Environment Applications Tel Aviv, Israel, March 16-18, 2009

Apan, Armando; Held, Alex; Phinn, Stuart; Markley, John Formulation and assessment of narrow-band vegetation indices from EO-1 hyperion imagery for discriminating sugarcane disease 2003 2003 Spatial Sciences Institute Conference: Spatial Knowledge Without Boundaries (SSC2003)

Bannari, A.; Morin, D.; Bonn, F.; Huete, A. R. A review of vegetation indices 1995 Remote Sensing Reviews

Barnes, E.M.; Clarke, T.R.; Richards, S.E.; Colaizzi, P.D.; Haberland, J.; Kostrzewski, M.; Waller, P.; Choi, C., Riley, E.; Thompson, T.; Lascano, R.J.; Li, H.; Moran, M.S. Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data 2000 Proc. 5th Int. Conf. Precis Agric

Barnes, J. D.; Balaguer, L.; Manrique, E.; Elvira, S.; Davison, A. W. A Reappraisal of the Use of Dmso for the Extraction and Determination of Chlorophylls-a and Chlorophylls-B in Lichens and Higher-Plants 1992 Environmental and Experimental Botany

Blackburn, G. A. Spectral indices for estimating photosynthetic pigment concentrations: A test using senescent tree leaves 1998 International Journal of Remote Sensing

Carter, Gregory A. Ratios of leaf reflectances in narrow wavebands as indicators of plant stress 1994 International Journal of Remote Sensing

Carter, Gregory A.; Cibula, William G.; Miller, Richard L. Narrow-band Reflectance Imagery Compared with ThermalImagery for Early Detection of Plant Stress 1996 Journal of Plant Physiology
Chappelle, E.W.; Kim, M.S.; McMurtrey, J.E. Ratio analysis of reflectance spectra (RARS): an algorithm for the remote estimation of the concentrations of chlorophyll a, chlorophyll b, and carotenoids in soybean leaves 1992 Remote Sensing of Environment
Chen, J.M. Evaluation of vegetation indices and a modified simple ratio for boreal applications 1996 Canadian Journal of Remote Sensing
Daughtry, C. S. T.; Walthall, C. L.; Kim, M. S.; de Colstoun, E. Brown; McMurtrey Iii, J. E. Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance 2000 Remote Sensing of Environment
Daughtry, C.; Hunt, E. R.; Walthall, C. L.; Gish, T. J.; Liang, Shunlin; Kramer, E.J. Assessing the Spatial Distribution of Plant Litter 2001 Proceedings of the Tenth JPL Airborne Earth Science Workshop
Dobrowski, S. Z., Pushnik, J. C., Zarco-Tejada, P. J., & Ustin, S. L. (2005). Simple reflectance indices track heat and water stress-induced changes in steady-state chlorophyll fluorescence at the canopy scale. Remote Sensing of Environment, 97(3), 403-414.
Gamon, J. A.; Peñuelas, J.; Field, C. B. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency 1992 Remote Sensing of Environment
Gamon, J. A.; Surfus, J. S. Assessing leaf pigment content and activity with a reflectometer 1999 New Phytologist
Gitelson, A. A.; Kaufman, Y. J.; Stark, R.; Rundquist, D. Novel algorithms for remote estimation of vegetation fraction 2002 Remote Sensing of Environment
Gitelson, A. A.; Merzlyak, M. N. Remote estimation of chlorophyll content in higher plant leaves 1997 International Journal of Remote Sensing
Gitelson, Anatoly A.; Kaufman, Yoram J.; Merzlyak, Mark N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS 1996 Remote Sensing of Environment
Gitelson, Anatoly A.; Keydan, Galina P.; Merzlyak, Mark N. Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves 2006 Geophys. Res. Lett.
Gitelson, Anatoly A.; Merzlyak, Mark N.; Zur, Y.; Stark, R.; Gritz, U. Non-destructive and remote sensing techniques for estimation of vegetation status 2001 Third European Conference on Precision Agriculture
Gitelson, Anatoly A.; Viña, Andrés; Arkebauer, Timothy J.; Rundquist, Donald C.; Keydan, Galina: Leavitt, Bryan Remote estimation of leaf area index and green leaf biomass in maize canopies 2003 Geophys. Res. Lett.
Goel, Narendra S.; Qin, Wenhan Influences of canopy architecture on relationships between various vegetation indices and LAI and Fpar: A computer simulation 1994

Remote Sensing Reviews
Guyot, G.; Baret, F.; Major, D. J. High spectral resolution: Determination of specral shifts between the red and the near infrared 1988 International Archives of Photogrammetry and Remote Sensing
Haboudane, Driss; Miller, John R.; Pattey, Elizabeth; Zarco-Tejada, Pablo J.; Strachan, Ian B. Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture 2004 Remote Sensing of Environment
Haboudane, Driss; Miller, John R.; Tremblay, Nicolas; Zarco-Tejada, Pablo J.; Dextraze, Louise Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture 2002 Remote Sensing of Environment
Hardinsky, M. A.; Lemas, V. The influence of soil salinity, growth form, and leaf moisture on the spectral reflectance of Spartina alternifolia canopies 1983 Photogrammetric Engineering and Remote Sensing
Hernández-Clemente, R., Navarro-Cerrillo, R. M., Suárez, L., Morales, F., & Zarco-Tejada, P. J. (2011). Assessing structural effects on PRI for stress detection in conifer forests. Remote Sensing of Environment, 115(9), 2360-2375.
Herrmann, I.; Karnieli, A.; Bonfil, D. J.; Cohen, Y.; Alchanatis, V. SWIR-based spectral indices for assessing nitrogen content in potato fields 2010 International Journal of Remote Sensing
Huete, A. R. A soil-adjusted vegetation index (SAVI) 1988 Remote Sensing of Environment
Huete, A. R.; Liu, H. Q.; Batchily, K.; van Leeuwen, W. A comparison of vegetation indices over a global set of TM images for EOS-MODIS 1997 Remote Sensing of Environment
Hunt Jr, E. Raymond; Rock, Barrett N. Detection of changes in leaf water content using Near- and Middle-Infrared reflectances 1989 Remote Sensing of Environment
Kaufman, Y. J.;Tanre, D. Atmospherically resistant vegetation index (ARVI) for EOS-MODIS 1992 Geoscience and Remote Sensing, IEEE Transactions on
Lichtenthaler, Hartmut K. Vegetation Stress: an Introduction to the Stress Concept in Plants 1996 Journal of Plant Physiology
Merzlyak, Mark N.; Gitelson, Anatoly A.; Chivkunova, Olga B.; Rakitin, Victor Y. U. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening 1999 Physiologia Plantarum
Penuelas J., Baret F., Filella I. Semi-empirical indices to assess carotenoids/chlorophyll a ratio from leaf spectral reflectance 1995 Photosynthetica
Penuelas, J.; Filella, I.; Biel, C.; Serrano, L.; Save, R. The reflectance at the 950□970 nm region as an indicator of plant water status 1993 International Journal of Remote Sensing

Peñuelas, J.; Gamon, J. A.; Fredeen, A. L.; Merino, J.; Field, C. B. Reflectance indices associated with physiological changes in nitrogen- and water-limited sunflower leaves 1994 Remote Sensing of Environment
Penuelas, Josep; Filella, Iolanda; Gamon, John A. Assessment of photosynthetic radiation-use efficiency with spectral reflectance 1995 New Phytologist
Pinty, B.; Verstraete, M. M. GEMI: a non-linear index to monitor global vegetation from satellites 1992 Plant Ecology
Rondeaux, Geneviève; Steven, Michael; Baret, Frédéric Optimization of soil-adjusted vegetation indices 1996 Remote Sensing of Environment
Roujean, Jean-Louis; Breon, François-Marie Estimating PAR absorbed by vegetation from bidirectional reflectance measurements 1995 Remote Sensing of Environment
Rouse, J.W., Jr.; Haas,R.H.; Schell, J.A.; Deering, D.W. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation 1973 Prog. Rep. RSC 1978-1
Sims, Daniel A.; Gamon, John A. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages 2002 Remote Sensing of Environment
Tucker, C. J.; Elgin Jr, J. H.; McMurtrey Iii, J. E.; Fan, C. J. Monitoring corn and soybean crop development with hand-held radiometer spectral data 1979 Remote Sensing of Environment
Vogelmann, J.E.; Rock, B.N.; Moss, D.M. Red Edge Spectral Measurements from Sugar Maple Leaves 1993 International Journal of Remote Sensing
Zarco-Tejada, P. J.; Miller, J. R.; Noland, T. L.; Mohammed, G. H.; Sampson, P. H. Scaling-up and model inversion methods with narrow-band optical indices for chlorophyll content estimation in closed forest canopies with hyperspectral data 2001 IEEE Transactions on Geoscience and Remote Sensing
Zarco-Tejada, Pablo J., Alberto Berjón, Raúl López-Lozano, John R. Miller, P. Martín, Victoria Cachorro, M. R. González, and A. De Frutos. "Assessing vineyard condition with hyperspectral indices: Leaf and canopy reflectance simulation in a row-structured discontinuous canopy." Remote Sensing of Environment 99, no. 3 (2005): 271-287.