Table of Contents

TERRA-REF Documentation

About this book

This book describes the TERRA-REF data collection, computing, and analysis pipelines.

- What data is available?
- Where do I get the data?
- User tutorials

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.

Versions

The first edition will be published in **November 2016**.

Data Sources

Field phenotyping

Maricopa Agricultural Center (MAC), Arizona

Over 300 sorghum accessions and recombinant inbred lines were planted at the University of Arizona Maricopa Agricultural Center and USDA Arid Land Research Station in Maricopa, Arizona. Field Scanner

- 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.
- The **PhenoTractor** is fitted with a sensor frame that supports a real time kinematic (RTK) satellite navigation antenna, a sonar transducer, an infrared temperature (IRT) scanner, and three GreenSeeker crop sensing systems.
- **UAV** (release V1)
- **Manually Collected Field Data** - Data will be collected manually to verify the sensor-collected data.

Kansas State University

- Tractor - coming 2017

- UAV - coming 2017

Controlled-environment phenotyping

Donald Danforth Plant Science Center, Missouri

The Bellwether Foundation Phenotyping Facility is a climate controlled 70 m^2 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.

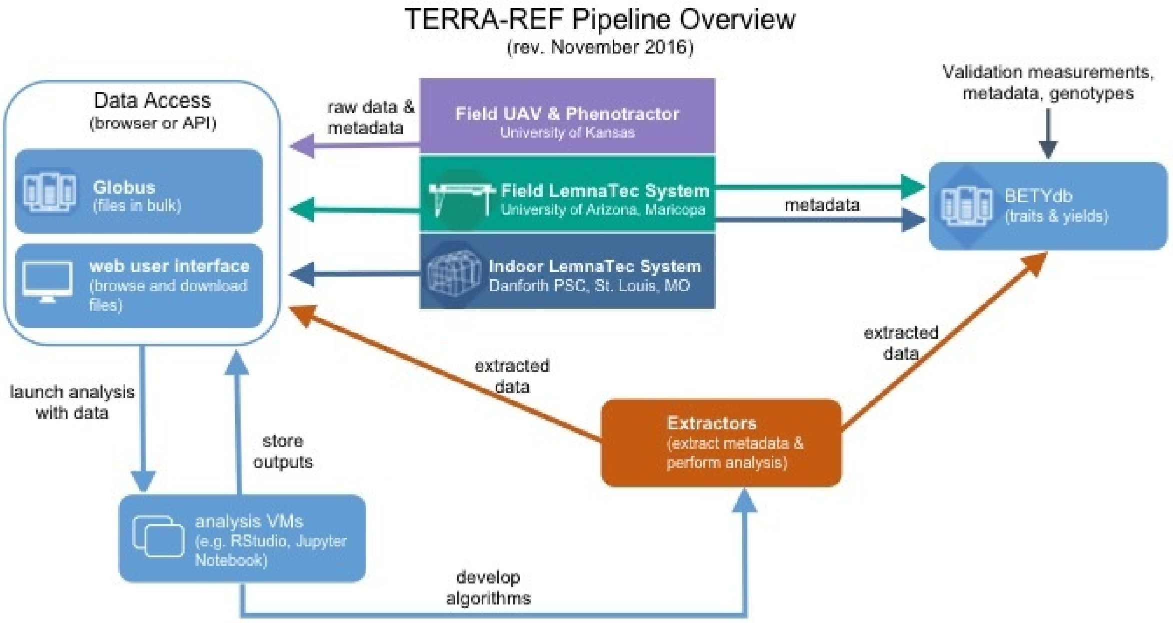
Genomics

HudsonAlpha Institute for Biotechnology. Alabama

coming 2017

Software

TERRA-REF uses the following software.



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. Data is available for Globus transfer via the Terraref endpoint. Use Globus Online when you want to transfer data from the TERRA-REF system for local analysis. For more information, see Using Globus.

BETYdb (phenotype data)

BETYdb contains the phenotype data with 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.

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.

Analysis Tools

The Clowder system supports launching external analysis environments such as RStudio and Jupyter Notebooks for anlaysis.

CoGe

CoGe contains genomic information and sequence data. For more information, see Using CoGe.

Scientific objectives and experimental design

Phenotyping

The TERRA-REF project is phenotyping the same genotypes of sorghum at multiple locations

- Automated Lemnatec Scanalyzer Field System at Maricopa Agricultural Center (MAC)
- PhenoTractors on parallel plots at MAC and Kansas State University (KSU)
- UAV platform on parallel plots at KSU
- Controlled-environment phenotyping systems at the Danforth Center
  - Manually collected field data at all locations

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.

Maricopa Agricultural Center (MAC), Arizona

Three hundred thirty one lines were planted in 2016. Plantings occurred both under and west of the gantry system.

- Experiments planned for 2016
- Field layouts under the gantry and west of the gantry in 2016.

Automated Field Scanner System

The Lemnatec Scanalyzer Field Scanner 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 system. Detailed information for each sensor including name, variable measured, and field of view are available here. The planned sensor missions and their objectives for 2016 are available here.

Phenotractor

The PhenoTractor at MAC is fitted with a sensor frame that supports a real time kinematic (RTK) satellite navigation antenna, a sonar transducer, an infrared temperature (IRT) scanner, and three GreenSeeker crop sensing systems.

UAV

Coming 2017

Manually collected field data

In progress

Automated controlled-environment phenotyping, Missouri

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.

Kansas State University

- **PhenoTractor** - Coming 2017
- **UAV** - Coming 2017
- **Manually collected field data** - Coming 2017

HudsonAlpha - Genomics

- Coming 2017

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**

A template for documenting protocols is available here.

Controled Environment Protocol

```
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
- 32^{o}C day/22^{o}C night temperature
- 60% relative humidity
- 700 umol/m^{2}/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

Manually Collected 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.
	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.

Lodging	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 will be determined according to Vanderlip (1993). Before heading, developmental stages were based on the

Phenology	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.
	Stomatal conductance was assessed using a leaf porometer

Transpiration/stomatal conductance	(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. http://dx.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_1V6QiKYLbIq1jIqXLJ84/edit#

PhenoTractor Sensor Protocols

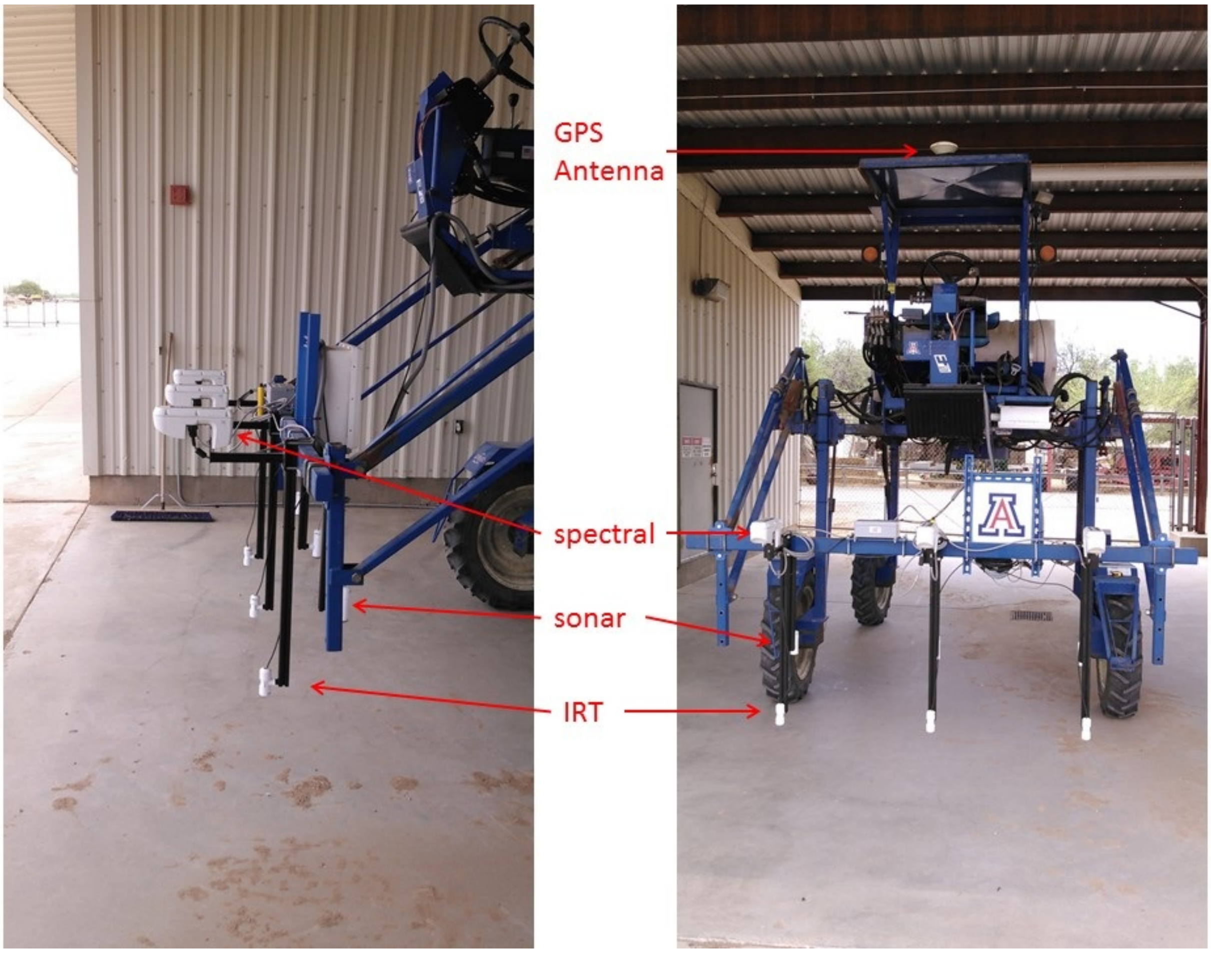
Authors: Matthew Maimaitiyiming and Wasit Wulamu
Center for Sustainability, Saint Louis University
January 8, 2017

Abstract

Materials

See Andrade-Sanchez et al 2014.

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



Picture of Phenotractor Sensors

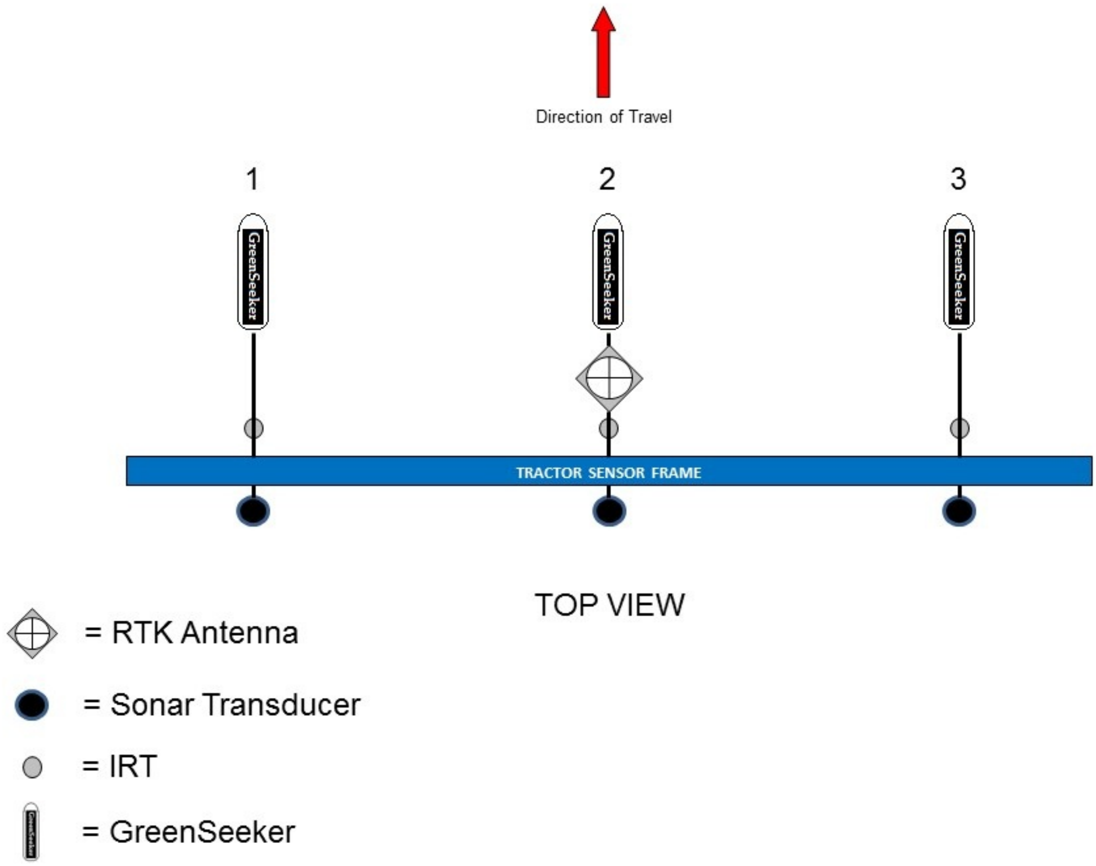


Diagram of sensor attachments

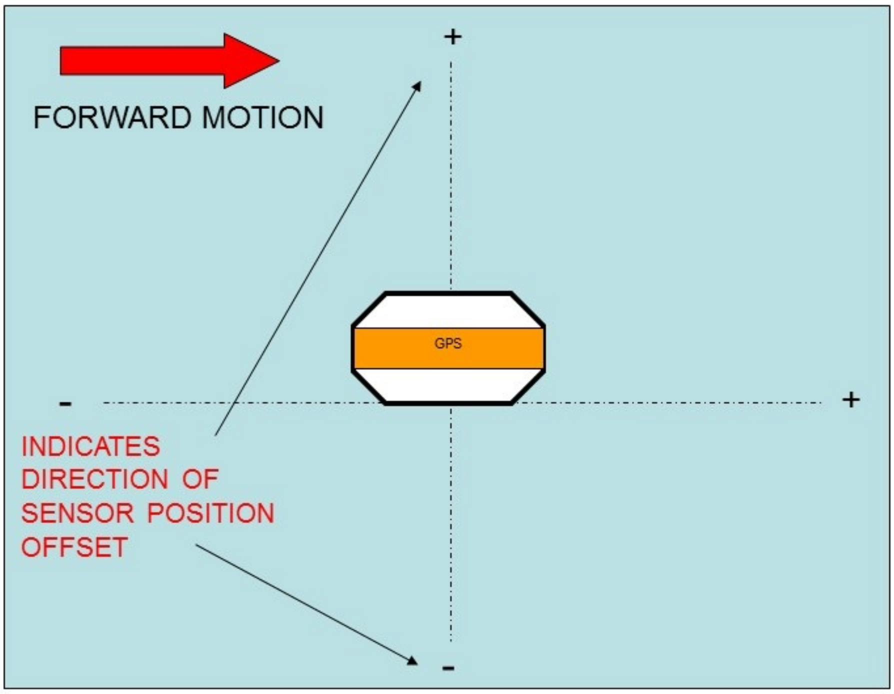


Diagram of Sensor Offset

Methods

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.

Phenotractor system configuration

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:

$$CH = E*rm{s} – E*\rm{ref} - S_\rm{output}$$

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. Infrared radiometer (IRT) sensors were used measure canopy temperature and temperature values were recoded as degree Celsius (°C).
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:

$$NDVI = (\frac{\rho*rm{NIR}-\rho*\rm{red}}{\rho*rm{NIR}+\rho*\rm{red}}$$

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

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\/0ByXIACImwxA7dVNHa3pTYkFjdWc\/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.

Template Protocol

```
Authors:
```

Abstract

Materials

Equipment

Procedure

Recipes

References

UAV Sensor Protocols

```
Authors: Rick Ward
```

Abstract

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

Experimental Design

Field phenotyping

- Maricopa Agricultural Center (MAC), Arizona
- Kansas State University - coming 2017

Controlled-environment phenotyping

- Donald Danforth Plant Science Center, Missouri

Genomics

- HudsonAlpha Institute for Biotechnology, Alabama - coming 2017

Controlled-Environment Phenotyping Experimental Designs

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

The Scanalyzer 3D platform 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)

Sorghum Lines Planted at Danforth (Year 1)

Experiment LT1A (TM015)

ATLAS LEOTI PI_144134 PI_145619 PI_145626 PI_145632 PI_145633 PI_146890 PI_147224 PI_152591 PI_152651 PI_152694 PI_152727 PI_152728 PI_152730 PI_152733 PI_152751 PI_152771 PI_152816 PI_152828 PI_152860 PI_152862 PI_152923 PI_152961 PI_152963 PI_152965 PI_152966 PI_152967 PI_152971 PI_153877 PI_154750 PI_154844 PI_154846 PI_154944 PI_154987 PI_154988 PI_155149 PI_155516 PI_155760 PI_155885 PI_156178 PI_156203 PI_156217 PI_156268 PI_156326 PI_156330 PI_156393 PI_156463 PI_156487 PI_156871 PI_156890 PI_157030 PI_157033 PI_157035 PI_157804 PI_167093 PI_170787 PI_175919 PI_176766 PI_179749 PI_180348 PI_181080 PI_181083 PI_195754 PI_196049 PI_196583 PI_196586 PI_196598 PI_197542 PI_19770 PI_213900 PI_217691 PI_218112 PI_221548 PI_221651 PI_226096 PI_22913 PI_229841 PI_251672 PI_253986 PI_255239 PI_255744 PI_257599 PI_257600 PI_266927 PI_267573 PI_273465 PI_273969 PI_276837 PI_297130 PI_297155 PI_297171 PI_302252 PI_303658 PI_329256 PI_329286 PI_329299 PI_329300 PI_329301 PI_329310 PI_329319 PI_329326 PI_329333 PI_329338 PI_329351 PI_329394 PI_329403 PI_329435 PI_329440 PI_329465 PI_329466 PI_329471 PI_329473 PI_329478 PI_329480 PI_329501 PI_329506 PI_329510 PI_329511 PI_329517 PI_329518 PI_329519 PI_329541 PI_329545 PI_329546 PI_329550 PI_329569 PI_329570 PI_329584 PI_329585 PI_329605 PI_329614 PI_329615 PI_329618 PI_329632 PI_329644 PI_329645 PI_329646 PI_329665 PI_329673 PI_329699 PI_329702 PI_329710 PI_329711 PI_329841 PI_329843 PI_329864 PI_329865 PI_330168 PI_330169 PI_330181 PI_330182 PI_330184 PI_330185 PI_330195 PI_330196 PI_330199 PI_330796 PI_330803 PI_330807 PI_330833 PI_330858 PI_337680 PI_337689 PI_35038 PI_365512 PI_452542 PI_452619 PI_452692 PI_453696 PI_455217 PI_455221 PI_455280 PI_455301 PI_455307 PI_505717 PI_505722 PI_505735 PI_506030 PI_506069 PI_506114 PI_506122 PI_508366 PI_510757 PI_511355 PI_513898 PI_514456 PI_521019 PI_521152 PI_521280

Experiment LT1B (TM016)

PI_521290 PI_524475 PI_525049 PI_52606 PI_526905 PI_527045 PI_533792 PI_533902 PI_533998 PI_534120 PI_534165 PI_535783 PI_535785 PI_535792 PI_535793 PI_535794 PI_535795 PI_535796 PI_540518 PI_542718 PI_550604 PI_561840 PI_562730 PI_562732 PI_562781 PI_562897 PI_562970 PI_562971 PI_562981 PI_562982 PI_562985 PI_562990 PI_562991 PI_562994 PI_562997 PI_562998 PI_563002 PI_563006 PI_563009 PI_563020 PI_563021 PI_563022 PI_563032 PI_563196 PI_563222 PI_563295 PI_563329 PI_563330

PI_563331 PI_563332 PI_563338 PI_563348 PI_563350 PI_563355 PI_564163 PI_566819 PI_568717 PI_569090 PI_569097 PI_569148 PI_569244 PI_569264 PI_569416 PI_569418 PI_569419 PI_569420 PI_569421 PI_569422 PI_569423 PI_569425 PI_569427 PI_569433 PI_569435 PI_569443 PI_569444 PI_569445 PI_569447 PI_569452 PI_569453 PI_569454 PI_569455 PI_569457 PI_569458 PI_569459 PI_569460 PI_569462 PI_569465 PI_569886 PI_570031 PI_570038 PI_570042 PI_570047 PI_570053 PI_570071 PI_570073 PI_570074 PI_570075 PI_570076 PI_570085 PI_570087 PI_570090 PI_570091 PI_570096 PI_570106 PI_570109 PI_570110 PI_570114 PI_570145 PI_570254 PI_570371 PI_570373 PI_570388 PI_570393 PI_570400 PI_570431 PI_573193 PI_576399 PI_576401 PI_583832 PI_585406 PI_585448 PI_585452 PI_585454 PI_585461 PI_585467 PI_585577 PI_585608 PI_585954 PI_585961 PI_585966 PI_586435 PI_586443 PI_586541 PI_593916 PI_619807 PI_619838 PI_620072 PI_620157 PI_63715 PI_641807 PI_641810 PI_641815 PI_641817 PI_641821 PI_641824 PI_641829 PI_641830 PI_641835 PI_641836 PI_641850 PI_641860 PI_641862 PI_641892 PI_641909 PI_642998 PI_643008 PI_643016 PI_646242 PI_646251 PI_646266 PI_651491 PI_651493 PI_651495 PI_651496 PI_651497 PI_653616 PI_653617 PI_655972 PI_655978 PI_655981 PI_655983 PI_656015 PI_656026 PI_656035 PI_656065 PI_92270 PI_329471 PI_329506 PI_329569 PI_337680 PI_452692 PI_455217 PI_152730 PI_329311 NTJ2 M81e CK60B B_Az9504 ICSV700 China 17

Coming soon

Sorghum Lines for Genomics Experiment (Year 1)

accession	taxid	organism common name	subspecific genetic lineage rank	subspecific genetic lineage name	ploidy	number of replicons
BTx623	4558	sorghum	cultivar	PI 564163	diploid	10
PI 144134	4558	sorghum	cultivar	PI 144134	diploid	10
PI 145619	4558	sorghum	cultivar	PI 145619	diploid	10
PI 145626	4558	sorghum	cultivar	PI 145626	diploid	10
PI 145632	4558	sorghum	cultivar	PI 145632	diploid	10
PI 145633	4558	sorghum	cultivar	PI 145633	diploid	10
PI 146890	4558	sorghum	cultivar	PI 146890	diploid	10
PI 152694	4558	sorghum	cultivar	PI 152694	diploid	10
PI 152728	4558	sorghum	cultivar	PI 152728	diploid	10

PI 152751	4558	sorghum	cultivar	PI 152751	diploid	10
PI 152771	4558	sorghum	cultivar	PI 152771	diploid	10
PI 152816	4558	sorghum	cultivar	PI 152816	diploid	10
PI 152828	4558	sorghum	cultivar	PI 152828	diploid	10
PI 152860	4558	sorghum	cultivar	PI 152860	diploid	10
PI 152862	4558	sorghum	cultivar	PI 152862	diploid	10
PI 152923	4558	sorghum	cultivar	PI 152923	diploid	10
PI 152961	4558	sorghum	cultivar	PI 152961	diploid	10
PI 152963	4558	sorghum	cultivar	PI 152963	diploid	10
PI 152965	4558	sorghum	cultivar	PI 152965	diploid	10
PI 152966	4558	sorghum	cultivar	PI 152966	diploid	10

PI 152967	4558	sorghum	cultivar	PI 152967	diploid	10
PI 152971	4558	sorghum	cultivar	PI 152971	diploid	10
PI 154944	4558	sorghum	cultivar	PI 154944	diploid	10
PI 155149	4558	sorghum	cultivar	PI 155149	diploid	10
PI 155516	4558	sorghum	cultivar	PI 155516	diploid	10
PI 155760	4558	sorghum	cultivar	PI 155760	diploid	10
PI 155885	4558	sorghum	cultivar	PI 155885	diploid	10
PI 156018	4558	sorghum	cultivar	PI 156018	diploid	10
PI 156217	4558	sorghum	cultivar	PI 156217	diploid	10
PI 156326	4558	sorghum	cultivar	PI 156326	diploid	10
PI 156463	4558	sorghum	cultivar	PI 156463	diploid	10
PI 156871	4558	sorghum	cultivar	PI 156871	diploid	10

PI 156890	4558	sorghum	cultivar	PI 156890	diploid	10
PI 157030	4558	sorghum	cultivar	PI 157030	diploid	10
PI 157033	4558	sorghum	cultivar	PI 157033	diploid	10
PI 157035	4558	sorghum	cultivar	PI 157035	diploid	10
PI 157804	4558	sorghum	cultivar	PI 157804	diploid	10
PI 167093	4558	sorghum	cultivar	PI 167093	diploid	10
PI 170787	4558	sorghum	cultivar	PI 170787	diploid	10
PI 180348	4558	sorghum	cultivar	PI 180348	diploid	10
PI 181080	4558	sorghum	cultivar	PI 181080	diploid	10
PI 181083	4558	sorghum	cultivar	PI 181083	diploid	10
PI 181899	4558	sorghum	cultivar	PI 181899	diploid	10

PI 196583	4558	sorghum	cultivar	PI 196583	diploid	10
PI 196586	4558	sorghum	cultivar	PI 196586	diploid	10
PI 196598	4558	sorghum	cultivar	PI 196598	diploid	10
PI 218112	4558	sorghum	cultivar	PI 218112	diploid	10
PI 221548	4558	sorghum	cultivar	PI 221548	diploid	10
PI 221651	4558	sorghum	cultivar	PI 221651	diploid	10
PI 250583	4558	sorghum	cultivar	PI 250583	diploid	10
PI 251672	4558	sorghum	cultivar	PI 251672	diploid	10
PI 253986	4558	sorghum	cultivar	PI 253986	diploid	10
PI 255744	4558	sorghum	cultivar	PI 255744	diploid	10
PI 257599	4558	sorghum	cultivar	PI 257599	diploid	10

PI 257600	4558	sorghum	cultivar	PI 257600	diploid	10
PI 266927	4558	sorghum	cultivar	PI 266927	diploid	10
PI 267573	4558	sorghum	cultivar	PI 267573	diploid	10
PI 273969	4558	sorghum	cultivar	PI 273969	diploid	10
PI 276837	4558	sorghum	cultivar	PI 276837	diploid	10
PI 291246	4558	sorghum	cultivar	PI 291246	diploid	10
PI 303658	4558	sorghum	cultivar	PI 303658	diploid	10
PI 329256	4558	sorghum	cultivar	PI 329256	diploid	10
PI 329286	4558	sorghum	cultivar	PI 329286	diploid	10
PI 329299	4558	sorghum	cultivar	PI 329299	diploid	10
PI 329310	4558	sorghum	cultivar	PI 329310	diploid	10

PI 329319	4558	sorghum	cultivar	PI 329319	diploid	10
PI 329326	4558	sorghum	cultivar	PI 329326	diploid	10
PI 329333	4558	sorghum	cultivar	PI 329333	diploid	10
PI 329338	4558	sorghum	cultivar	PI 329338	diploid	10
PI 329351	4558	sorghum	cultivar	PI 329351	diploid	10
PI 329394	4558	sorghum	cultivar	PI 329394	diploid	10
PI 329440	4558	sorghum	cultivar	PI 329440	diploid	10
PI 329465	4558	sorghum	cultivar	PI 329465	diploid	10
PI 329471	4558	sorghum	cultivar	PI 329471	diploid	10
PI 329480	4558	sorghum	cultivar	PI 329480	diploid	10
PI 329506	4558	sorghum	cultivar	PI 329506	diploid	10
PI 329510	4558	sorghum	cultivar	PI 329510	diploid	10

PI 329510	4558	sorghum	cultivar	PI 329510	diploid	10
PI 329511	4558	sorghum	cultivar	PI 329511	diploid	10
PI 329517	4558	sorghum	cultivar	PI 329517	diploid	10
PI 329518	4558	sorghum	cultivar	PI 329518	diploid	10
PI 329519	4558	sorghum	cultivar	PI 329519	diploid	10
PI 329569	4558	sorghum	cultivar	PI 329569	diploid	10
PI 329570	4558	sorghum	cultivar	PI 329570	diploid	10
PI 329584	4558	sorghum	cultivar	PI 329584	diploid	10
PI 329585	4558	sorghum	cultivar	PI 329585	diploid	10
PI 329605	4558	sorghum	cultivar	PI 329605	diploid	10
PI 365512	4558	sorghum	cultivar	PI 365512	diploid	10
PI 452542	4558	sorghum	cultivar	PI 452542	diploid	10

PI 452544	4558	sorghum	cultivar	PI 452544	diploid	10
PI 452619	4558	sorghum	cultivar	PI 452619	diploid	10
PI 452692	4558	sorghum	cultivar	PI 452692	diploid	10
PI 455217	4558	sorghum	cultivar	PI 455217	diploid	10
PI 455301	4558	sorghum	cultivar	PI 455301	diploid	10
PI 455307	4558	sorghum	cultivar	PI 455307	diploid	10
PI 506030	4558	sorghum	cultivar	PI 506030	diploid	10
PI 511355	4558	sorghum	cultivar	PI 511355	diploid	10
PI 513898	4558	sorghum	cultivar	PI 513898	diploid	10
PI 514456	4558	sorghum	cultivar	PI 514456	diploid	10
PI 521019	4558	sorghum	cultivar	PI 521019	diploid	10

PI 521280	4558	sorghum	cultivar	PI 521280	diploid	10
PI 521281	4558	sorghum	cultivar	PI 521281	diploid	10
PI 521290	4558	sorghum	cultivar	PI 521290	diploid	10
PI 524475	4558	sorghum	cultivar	PI 524475	diploid	10
PI 526905	4558	sorghum	cultivar	PI 526905	diploid	10
PI 527045	4558	sorghum	cultivar	PI 527045	diploid	10
PI 533902	4558	sorghum	cultivar	PI 533902	diploid	10
PI 533998	4558	sorghum	cultivar	PI 533998	diploid	10
PI 534047	4558	sorghum	cultivar	PI 534047	diploid	10
PI 534120	4558	sorghum	cultivar	PI 534120	diploid	10
PI 534165	4558	sorghum	cultivar	PI 534165	diploid	10
PI 535783	4558	sorghum	cultivar	PI 535783	diploid	10

PI 535785	4558	sorghum	cultivar	PI 535785	diploid	10
PI 535792	4558	sorghum	cultivar	PI 535792	diploid	10
PI 535793	4558	sorghum	cultivar	PI 535793	diploid	10
PI 535794	4558	sorghum	cultivar	PI 535794	diploid	10
PI 535795	4558	sorghum	cultivar	PI 535795	diploid	10
PI 535796	4558	sorghum	cultivar	PI 535796	diploid	10
PI 540518	4558	sorghum	cultivar	PI 540518	diploid	10
PI 550604	4558	sorghum	cultivar	PI 550604	diploid	10
PI 552851	4558	sorghum	cultivar	PI 552851	diploid	10
PI 561072	4558	sorghum	cultivar	PI 561072	diploid	10
PI 562717	4558	sorghum	cultivar	PI 562717	diploid	10

PI 562732	4558	sorghum	cultivar	PI 562732	diploid	10
PI 562781	4558	sorghum	cultivar	PI 562781	diploid	10
PI 562897	4558	sorghum	cultivar	PI 562897	diploid	10
PI 562970	4558	sorghum	cultivar	PI 562970	diploid	10
PI 562971	4558	sorghum	cultivar	PI 562971	diploid	10
PI 562981	4558	sorghum	cultivar	PI 562981	diploid	10
PI 562982	4558	sorghum	cultivar	PI 562982	diploid	10
PI 562985	4558	sorghum	cultivar	PI 562985	diploid	10
PI 562990	4558	sorghum	cultivar	PI 562990	diploid	10
PI 562991	4558	sorghum	cultivar	PI 562991	diploid	10
PI 562997	4558	sorghum	cultivar	PI 562997	diploid	10

PI 563002	4558	sorghum	cultivar	PI 563002	diploid	10
PI 563009	4558	sorghum	cultivar	PI 563009	diploid	10
PI 563020	4558	sorghum	cultivar	PI 563020	diploid	10
PI 563021	4558	sorghum	cultivar	PI 563021	diploid	10
PI 563022	4558	sorghum	cultivar	PI 563022	diploid	10
PI 563032	4558	sorghum	cultivar	PI 563032	diploid	10
PI 563196	4558	sorghum	cultivar	PI 563196	diploid	10
PI 563355	4558	sorghum	cultivar	PI 563355	diploid	10
PI 564163	4558	sorghum	cultivar	PI 564163	diploid	10
PI 570073	4558	sorghum	cultivar	PI 570073	diploid	10
PI 570074	4558	sorghum	cultivar	PI 570074	diploid	10

PI 570074	4558	sorghum	cultivar	PI 570074	diploid	10
PI 570076	4558	sorghum	cultivar	PI 570076	diploid	10
PI 570085	4558	sorghum	cultivar	PI 570085	diploid	10
PI 570087	4558	sorghum	cultivar	PI 570087	diploid	10
PI 570090	4558	sorghum	cultivar	PI 570090	diploid	10
PI 570091	4558	sorghum	cultivar	PI 570091	diploid	10
PI 570096	4558	sorghum	cultivar	PI 570096	diploid	10
PI 570106	4558	sorghum	cultivar	PI 570106	diploid	10
PI 570110	4558	sorghum	cultivar	PI 570110	diploid	10
PI 570114	4558	sorghum	cultivar	PI 570114	diploid	10
PI 570145	4558	sorghum	cultivar	PI 570145	diploid	10
PI 570371	4558	sorghum	cultivar	PI 570371	diploid	10

PI 570393	4558	sorghum	cultivar	PI 570393	diploid	10
PI 570400	4558	sorghum	cultivar	PI 570400	diploid	10
PI 573193	4558	sorghum	cultivar	PI 573193	diploid	10
PI 576399	4558	sorghum	cultivar	PI 576399	diploid	10
PI 583832	4558	sorghum	cultivar	PI 583832	diploid	10
PI 585406	4558	sorghum	cultivar	PI 585406	diploid	10
PI 585448	4558	sorghum	cultivar	PI 585448	diploid	10
PI 585452	4558	sorghum	cultivar	PI 585452	diploid	10
PI 585454	4558	sorghum	cultivar	PI 585454	diploid	10
PI 585461	4558	sorghum	cultivar	PI 585461	diploid	10
PI 585467	4558	sorghum	cultivar	PI 585467	diploid	10
PI 585577	4558	sorghum	cultivar	PI 585577	diploid	10

PI 585608	4558	sorghum	cultivar	PI 585608	diploid	10
PI 585961	4558	sorghum	cultivar	PI 585961	diploid	10
PI 585966	4558	sorghum	cultivar	PI 585966	diploid	10
PI 586541	4558	sorghum	cultivar	PI 586541	diploid	10
PI 593916	4558	sorghum	cultivar	PI 593916	diploid	10
PI 619807	4558	sorghum	cultivar	PI 619807	diploid	10
PI 619838	4558	sorghum	cultivar	PI 619838	diploid	10
PI 620072	4558	sorghum	cultivar	PI 620072	diploid	10
PI 641824	4558	sorghum	cultivar	PI 641824	diploid	10
PI 641834	4558	sorghum	cultivar	PI 641834	diploid	10
PI 651493	4558	sorghum	cultivar	PI 651493	diploid	10

PI 653616	4558	sorghum	cultivar	PI 653616	diploid	10
PI 653617	4558	sorghum	cultivar	PI 653617	diploid	10
PI 655978	4558	sorghum	cultivar	PI 655978	diploid	10
PI 655981	4558	sorghum	cultivar	PI 655981	diploid	10
PI 656015	4558	sorghum	cultivar	PI 656015	diploid	10
PI 656026	4558	sorghum	cultivar	PI 656026	diploid	10
PI 656035	4558	sorghum	cultivar	PI 656035	diploid	10
PI 656056	4558	sorghum	cultivar	PI 656056	diploid	10
PI 656065	4558	sorghum	cultivar	PI 656065	diploid	10
PI 92270	4558	sorghum	cultivar	PI 92270	diploid	10
rio	4558	sorghum	cultivar	PI 563295	diploid	10

Sorghum Lines for Genomics Experiment (Year 2)

accession	taxid	organism common name	subspecific genetic lineage rank	subspecific genetic lineage name	ploidy	number of replicons
PI 300118	4558	sorghum	cultivar	PI 300118	diploid	10
PI 300119	4558	sorghum	cultivar	PI 300119	diploid	10
PI 569425	4558	sorghum	cultivar	PI 569425	diploid	10
PI 651491	4558	sorghum	cultivar	PI 651491	diploid	10
PI 569443	4558	sorghum	cultivar	PI 569443	diploid	10
PI 569433	4558	sorghum	cultivar	PI 569433	diploid	10
PI 19770	4558	sorghum	cultivar	PI 19770	diploid	10
PI 641860	4558	sorghum	cultivar	PI 641860	diploid	10
PI 651495	4558	sorghum	cultivar	PI 651495	diploid	10

PI 566819	4558	sorghum	cultivar	PI 566819	diploid	10
PI 569446	4558	sorghum	cultivar	PI 569446	diploid	10
PI 453106	4558	sorghum	cultivar	PI 453106	diploid	10
PI 453177	4558	sorghum	cultivar	PI 453177	diploid	10
PI 453696	4558	sorghum	cultivar	PI 453696	diploid	10
PI 455280	4558	sorghum	cultivar	PI 455280	diploid	10
PI 152651	4558	sorghum	cultivar	PI 152651	diploid	10
PI 641821	4558	sorghum	cultivar	PI 641821	diploid	10
PI 563329	4558	sorghum	cultivar	PI 563329	diploid	10
PI 330858	4558	sorghum	cultivar	PI 330858	diploid	10
PI 563332	4558	sorghum	cultivar	PI 563332	diploid	10

PI 563338	4558	sorghum	cultivar	PI 563338	diploid	10
PI 563350	4558	sorghum	cultivar	PI 563350	diploid	10
PI 329583	4558	sorghum	cultivar	PI 329583	diploid	10
PI 329615	4558	sorghum	cultivar	PI 329615	diploid	10
PI 329617	4558	sorghum	cultivar	PI 329617	diploid	10
PI 329644	4558	sorghum	cultivar	PI 329644	diploid	10
PI 329645	4558	sorghum	cultivar	PI 329645	diploid	10
PI 329646	4558	sorghum	cultivar	PI 329646	diploid	10
PI 329673	4558	sorghum	cultivar	PI 329673	diploid	10
PI 329841	4558	sorghum	cultivar	PI 329841	diploid	10
PI 329843	4558	sorghum	cultivar	PI 329843	diploid	10

PI 329865	4558	sorghum	cultivar	PI 329865	diploid	10
PI 330167	4558	sorghum	cultivar	PI 330167	diploid	10
PI 330169	4558	sorghum	cultivar	PI 330169	diploid	10
PI 330181	4558	sorghum	cultivar	PI 330181	diploid	10
PI 330182	4558	sorghum	cultivar	PI 330182	diploid	10
PI 330184	4558	sorghum	cultivar	PI 330184	diploid	10
PI 330196	4558	sorghum	cultivar	PI 330196	diploid	10
PI 330199	4558	sorghum	cultivar	PI 330199	diploid	10
PI 330796	4558	sorghum	cultivar	PI 330796	diploid	10
PI 330803	4558	sorghum	cultivar	PI 330803	diploid	10
PI 337680	4558	sorghum	cultivar	PI 337680	diploid	10

PI 337689	4558	sorghum	cultivar	PI 337689	diploid	10
PI 175919	4558	sorghum	cultivar	PI 175919	diploid	10
PI 569097	4558	sorghum	cultivar	PI 569097	diploid	10
PI 569148	4558	sorghum	cultivar	PI 569148	diploid	10
PI 569886	4558	sorghum	cultivar	PI 569886	diploid	10
PI 586499	4558	sorghum	cultivar	PI 586499	diploid	10
PI 179749	4558	sorghum	cultivar	PI 179749	diploid	10
PI 156330	4558	sorghum	cultivar	PI 156330	diploid	10
PI 569435	4558	sorghum	cultivar	PI 569435	diploid	10
PI 569465	4558	sorghum	cultivar	PI 569465	diploid	10
PI 569453	4558	sorghum	cultivar	PI 569453	diploid	10

PI 569423	4558	sorghum	cultivar	PI 569423	diploid	10
PI 569455	4558	sorghum	cultivar	PI 569455	diploid	10
PI 569459	4558	sorghum	cultivar	PI 569459	diploid	10
PI 569427	4558	sorghum	cultivar	PI 569427	diploid	10
PI 152733	4558	sorghum	cultivar	PI 152733	diploid	10
PI 156178	4558	sorghum	cultivar	PI 156178	diploid	10
PI 176766	4558	sorghum	cultivar	PI 176766	diploid	10
PI 569457	4558	sorghum	cultivar	PI 569457	diploid	10
PI 569445	4558	sorghum	cultivar	PI 569445	diploid	10
PI 568717	4558	sorghum	cultivar	PI 568717	diploid	10
PI 153877	4558	sorghum	cultivar	PI 153877	diploid	10
PI 525049	4558	sorghum	cultivar	PI 525049	diploid	10

PI 525049	4558	sorghum	cultivar	PI 525049	diploid	10
PI 152727	4558	sorghum	cultivar	PI 152727	diploid	10
PI 329435	4558	sorghum	cultivar	PI 329435	diploid	10
PI 576401	4558	sorghum	cultivar	PI 576401	diploid	10
PI 217691	4558	sorghum	cultivar	PI 217691	diploid	10
PI 569447	4558	sorghum	cultivar	PI 569447	diploid	10
PI 533792	4558	sorghum	cultivar	PI 533792	diploid	10
PI 521152	4558	sorghum	cultivar	PI 521152	diploid	10
PI 542718	4558	sorghum	cultivar	PI 542718	diploid	10
PI 641892	4558	sorghum	cultivar	PI 641892	diploid	10
PI 197542	4558	sorghum	cultivar	PI 197542	diploid	10

PI 651497	4558	sorghum	cultivar	PI 651497	diploid	10
PI 641850	4558	sorghum	cultivar	PI 641850	diploid	10
PI 185573	4558	sorghum	cultivar	PI 185573	diploid	10
PI 533752	4558	sorghum	cultivar	PI 533752	diploid	10
PI 213901	4558	sorghum	cultivar	PI 213901	diploid	10
PI 302221	4558	sorghum	cultivar	PI 302221	diploid	10
PI 656020	4558	sorghum	cultivar	PI 656020	diploid	10
PI 655986	4558	sorghum	cultivar	PI 655986	diploid	10
PI 655988	4558	sorghum	cultivar	PI 655988	diploid	10
PI 655995	4558	sorghum	cultivar	PI 655995	diploid	10
PI 329573	4558	sorghum	cultivar	PI 329573	diploid	10

PI 330122	4558	sorghum	cultivar	PI 330122	diploid	10
PI 576366	4558	sorghum	cultivar	PI 576366	diploid	10
PI 656051	4558	sorghum	cultivar	PI 656051	diploid	10
PI 653411	4558	sorghum	cultivar	PI 653411	diploid	10
PI 533759	4558	sorghum	cultivar	PI 533759	diploid	10
PI 534157	4558	sorghum	cultivar	PI 534157	diploid	10
PI 533964	4558	sorghum	cultivar	PI 533964	diploid	10
PI 606706	4558	sorghum	cultivar	PI 606706	diploid	10
PI 152591	4558	sorghum	cultivar	PI 152591	diploid	10
PI 147224	4558	sorghum	cultivar	PI 147224	diploid	10
PI 641862	4558	sorghum	cultivar	PI 641862	diploid	10

PI 641810	4558	sorghum	cultivar	PI 641810	diploid	10
PI 569416	4558	sorghum	cultivar	PI 569416	diploid	10
PI 569419	4558	sorghum	cultivar	PI 569419	diploid	10
PI 569454	4558	sorghum	cultivar	PI 569454	diploid	10
PI 641815	4558	sorghum	cultivar	PI 641815	diploid	10
PI 641817	4558	sorghum	cultivar	PI 641817	diploid	10
PI 569420	4558	sorghum	cultivar	PI 569420	diploid	10
PI 563348	4558	sorghum	cultivar	PI 563348	diploid	10
PI 329300	4558	sorghum	cultivar	PI 329300	diploid	10
PI 329301	4558	sorghum	cultivar	PI 329301	diploid	10
PI 329403	4558	sorghum	cultivar	PI 329403	diploid	10

PI 329466	4558	sorghum	cultivar	PI 329466	diploid	10
PI 329501	4558	sorghum	cultivar	PI 329501	diploid	10
PI 329545	4558	sorghum	cultivar	PI 329545	diploid	10
PI 329546	4558	sorghum	cultivar	PI 329546	diploid	10
PI 329550	4558	sorghum	cultivar	PI 329550	diploid	10
PI 329551	4558	sorghum	cultivar	PI 329551	diploid	10
PI 329554	4558	sorghum	cultivar	PI 329554	diploid	10
PI 329614	4558	sorghum	cultivar	PI 329614	diploid	10
PI 329699	4558	sorghum	cultivar	PI 329699	diploid	10
PI 329702	4558	sorghum	cultivar	PI 329702	diploid	10
PI 329710	4558	sorghum	cultivar	PI 329710	diploid	10

PI 330168	4558	sorghum	cultivar	PI 330168	diploid	10
PI 330185	4558	sorghum	cultivar	PI 330185	diploid	10
PI 330195	4558	sorghum	cultivar	PI 330195	diploid	10
PI 330833	4558	sorghum	cultivar	PI 330833	diploid	10
PI 302252	4558	sorghum	cultivar	PI 302252	diploid	10
PI 569244	4558	sorghum	cultivar	PI 569244	diploid	10
PI 569264	4558	sorghum	cultivar	PI 569264	diploid	10
PI 570031	4558	sorghum	cultivar	PI 570031	diploid	10
PI 570038	4558	sorghum	cultivar	PI 570038	diploid	10
PI 570039	4558	sorghum	cultivar	PI 570039	diploid	10
PI 570042	4558	sorghum	cultivar	PI 570042	diploid	10

PI 570053	4558	sorghum	cultivar	PI 570053	diploid	10
PI 570071	4558	sorghum	cultivar	PI 570071	diploid	10
PI 570431	4558	sorghum	cultivar	PI 570431	diploid	10
PI 646242	4558	sorghum	cultivar	PI 646242	diploid	10
PI 646251	4558	sorghum	cultivar	PI 646251	diploid	10
PI 646266	4558	sorghum	cultivar	PI 646266	diploid	10
PI 620157	4558	sorghum	cultivar	PI 620157	diploid	10
PI 154846	4558	sorghum	cultivar	PI 154846	diploid	10
PI 569418	4558	sorghum	cultivar	PI 569418	diploid	10
PI 641829	4558	sorghum	cultivar	PI 641829	diploid	10
PI 641830	4558	sorghum	cultivar	PI 641830	diploid	10

PI 156393	4558	sorghum	cultivar	PI 156393	diploid	10
PI 156487	4558	sorghum	cultivar	PI 156487	diploid	10
PI 643008	4558	sorghum	cultivar	PI 643008	diploid	10
PI 643016	4558	sorghum	cultivar	PI 643016	diploid	10
PI 569458	4558	sorghum	cultivar	PI 569458	diploid	10
PI 569422	4558	sorghum	cultivar	PI 569422	diploid	10
PI 569460	4558	sorghum	cultivar	PI 569460	diploid	10
PI 641835	4558	sorghum	cultivar	PI 641835	diploid	10
PI 641909	4558	sorghum	cultivar	PI 641909	diploid	10
PI 154988	4558	sorghum	cultivar	PI 154988	diploid	10
PI 586435	4558	sorghum	cultivar	PI 586435	diploid	10
PI 63715	4558	sorghum	cultivar	PI 63715	diploid	10

PI 63715	4558	sorghum	cultivar	PI 63715	diploid	10
PI 365512	4558	sorghum	cultivar	PI 365512	diploid	10
PI 585966	4558	sorghum	cultivar	PI 585966	diploid	10
PI 455307	4558	sorghum	cultivar	PI 455307	diploid	10
PI 562971	4558	sorghum	cultivar	PI 562971	diploid	10
PI 562897	4558	sorghum	cultivar	PI 562897	diploid	10
PI 562991	4558	sorghum	cultivar	PI 562991	diploid	10
PI 562994	4558	sorghum	cultivar	PI 562994	diploid	10
PI 329478	4558	sorghum	cultivar	PI 329478	diploid	10
PI 329518	4558	sorghum	cultivar	PI 329518	diploid	10
PI 570075	4558	sorghum	cultivar	PI 570075	diploid	10
PI 570110	4558	sorghum	cultivar	PI 570110	diploid	10

PI 152728	4558	sorghum	cultivar	PI 152728	diploid	10
PI 156203	4558	sorghum	cultivar	PI 156203	diploid	10
PI 570373	4558	sorghum	cultivar	PI 570373	diploid	10
PI 656030	4558	sorghum	cultivar	PI 656030	diploid	10
PI 533863	4558	sorghum	cultivar	PI 533863	diploid	10
PI 239439	4558	sorghum	cultivar	PI 239439	diploid	10
PI 629059	4558	sorghum	cultivar	PI 629059	diploid	10
PI 474825	4558	sorghum	cultivar	PI 474825	diploid	10
PI 656044	4558	sorghum	cultivar	PI 656044	diploid	10
PI 52606	4558	sorghum	cultivar	PI 52606	diploid	10
PI 655996	4558	sorghum	cultivar	PI 655996	diploid	10
PI 595741	4558	sorghum	cultivar	PI 595741	diploid	10

PI 595741	4558	sorghum	cultivar	PI 595741	diploid	10
PI 595699	4558	sorghum	cultivar	PI 595699	diploid	10
PI 656023	4558	sorghum	cultivar	PI 656023	diploid	10
PI 152971	4558	sorghum	cultivar	PI 152971	diploid	10
PI 570388	4558	sorghum	cultivar	PI 570388	diploid	10
PI 180348	4558	sorghum	cultivar	PI 180348	diploid	10
PI 535793	4558	sorghum	cultivar	PI 535793	diploid	10
PI 535783	4558	sorghum	cultivar	PI 535783	diploid	10
PI 257600	4558	sorghum	cultivar	PI 257600	diploid	10
PI 152751	4558	sorghum	cultivar	PI 152751	diploid	10
PI 154987	4558	sorghum	cultivar	PI 154987	diploid	10
PI 586541	4558	sorghum	cultivar	PI 586541	diploid	10

PI 152816	4558	sorghum	cultivar	PI 152816	diploid	10
PI 653616	4558	sorghum	cultivar	PI 653616	diploid	10

Maricopa Agricultural Center Experimental Design

Season 1 sorghum (April - July 2016) Season 2 sorghum (August - November 2016) Durum wheat (January 2017 -

Season 1 soghum (April - July 2016)

Three hundred thirty one lines were planted in Season 1.

Planting maps

- Under the scanner system
- West 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

West of scanner system

Experiment	Reps	Treatments	Experimental design
BAP	1	30 lines (12 PS, 12 sweet, 6 grain)	None - single rep planted for observation
RILs	3	60 RILs	Incomplete Block (row-column alpha lattice design)

Automated Phenotpying

The Lemnatec Scanalyzer Field Gantry 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 available here. The planned sensor missions and their objectives for 2016 are available here.

Manually Collected Field Data

emergence vigor emergence final stand counts plant heights node and tiller counts on marked plants phenology growth stage data leaf desiccation ratings radiation interception managements Incomplete harvest yield data

Season 2 soghum (August - November 2016)

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

Durum wheat

Experimental design

Automatically and manually collected field data plan

User Manual

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 Lemnatec field and greenhouse systems, UAVs and tractors
- Manually-collected fieldbooks and associated protocols
- Derived data, including phenomics data, from computational approaches
- Genomic pipeline data

Audience

The TERRA-REF reference dataset may be of interest to a variety of research communities including:

- Computer vision\/remote sensing\/image analysis (raw sensor data and metadata)
- Physiologists (plants and how they are growing)
- Robotics (gantry\/location\/orientation)
- Breeders (derived traits)
- Genomics\/bioinformaticians (genomics data)

What Data is Available

Real-time sensor data transfer by file number and size can be viewed here.

See Data Products for more information about individual data products and How to Access Data for instructions to access the data products.

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 imaging 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

- What data is available?
- Sensor calibration
- Fieldbooks and Protocols
- Data standards
- Geospatial information

Environmental conditions data

Environment conditions data is collected using the Vaisala CO_2, Thies Clima weather sensors as well as lightning, irrigation, and weather data collected at the Maricopa site.

Data access

Data is available via Clowder and Globus:

- **Clowder**:
    - co2sensor collection (Viasala CO_2)
    - EnvironmentalLogger (Thies Clima)
    - irrigation, lightning and weather collections (MAC Weather Station)
- **Globus**:
    - `/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 CO_2 Sensor collection
    - Thies Clima Sensor collection

Computational pipeline

Environmental Logger

- **Description:** EnvironmentalLogger raw files are converted to netCDF.
- **Output**: `/ua-mac/Level_1/EnvironmentLogger`

See also

- Geospatial information

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

Genomics Data

Genome resequencing data is available for 384 accessions of sorghum from the Bioenergy Association Panel (BAP). The available accessions from year 1 are listed here and from year 2 here. The raw and processed data is available using Globus.

Raw data (Bzip2 FASTQ):

```
/sites/hudson-alpha/raw_data/year1/
/sites/hudson-alpha/raw_data/year2/
```

Processed data (Gzipped VCF):

```
/sites/hudson-alpha/derived_data/year1/
/sites/hudson-alpha/derived_data/year2/
```

The output data are in variant call format (VCF), which contains single-nucleotide polymorphism (SNP) and insertion-deletion (indel) variation relative to the reference *Sorghum bicolor* v3.1 genome.

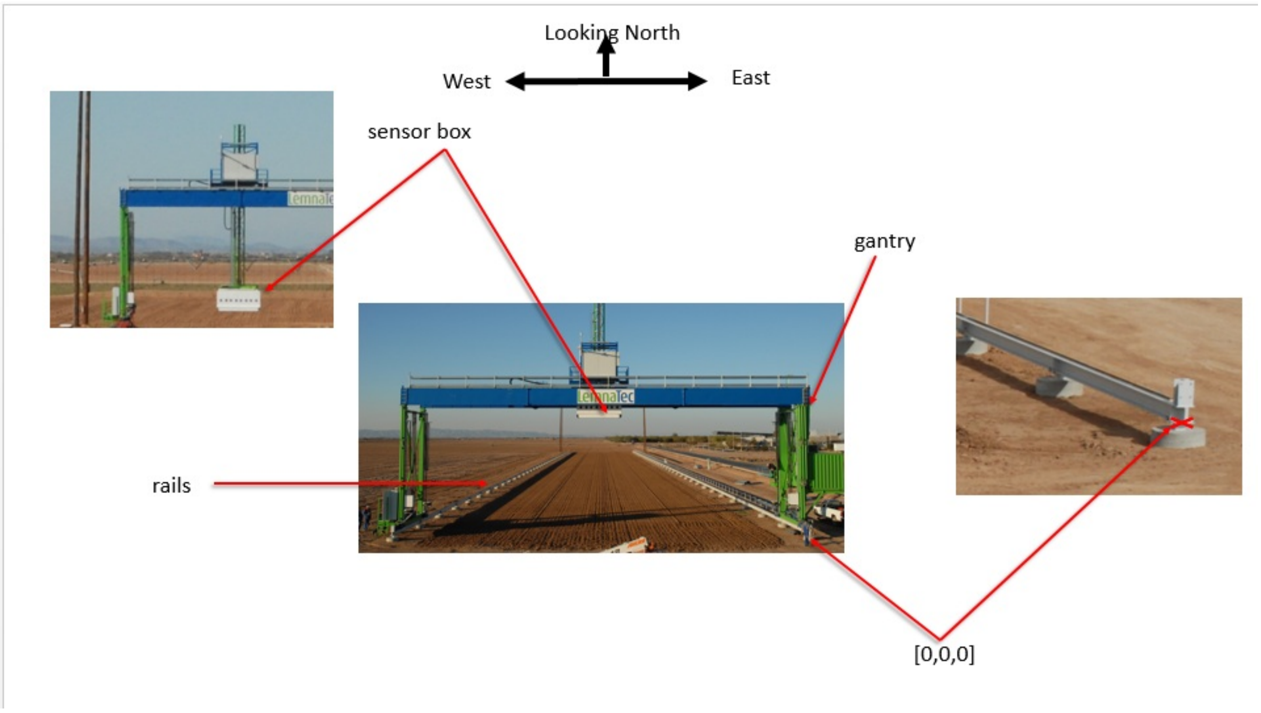
Geospatial information

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

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, perpendicular to g
round",
      "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, facing groun
d",
      "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, facing groun
d",
      "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 row and each range of the two planting seasons is available in the field book. The scanalyzer coordinates of each plot are transformed into the (EPSG:4326) USDA coordinates using the equations above. After that, a polygon of each plot can be generated using ST_GeomFromText funtion and inserted into the BETYdb through SQL statements.

An Rcode is available for generating SQL statements based on the scanalyzer coordinates of each plot, which takes range.csv and row.csv as standard inputs.

The range.csv should be in the following format:

range	x_south	x_north
1	...	...
2	...	...
3	...	...
...	...	...

And the row.csv should look like:

row	y_west	y_east
1	...	...
2	...	...
3	...	...
...	...	...

The output will be something look like:

```
INSERT INTO sites (sitename, geometry) VALUES ( 'MAC Field Scanner Field Plot 1 Season
 2',
ST_GeomFromText('POLYGON((-111.975049874375 33.0745312921391 353, -111.975033517034 33
.0745313124814 353,
-111.975033529346 33.0745670737771 353, -111.975049886694 33.0745670534346 353, -111.9
75049874375
33.0745312921391 353))', 4326));
```

Hyperspectral imaging data

Summary

Hyperspectral imaging data is collected using the Headwall VNIR and SWIR sensors.

Raw data access

Hyperspectral data is available via Clowder and Globus:

- **Clowder**:
  - SWIR Collection
  - VNIR Collection
- **Globus**:
  - `/ua-mac/raw_data/SWIR`
  - `/ua-mac/raw_data/VNIR`
- **Sensor information**:
  - Headwall SWIR
  - Headwall VNIR

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

Computational pipeline

Hyperspectral extractor

- **Description**: Processes HDF files into netCDF
- **Output**: `/sites/ua_mac/Level_1/hyperspectral`

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.

Raw data access

Thermal imaging data is available via Clowder and Globus:

- **Clowder**: flirIrCamera collection
- **Globus**: `/ua-mac/raw_data/flirIrCamera`
- **Sensor information**: FLIR Thermal Camera collection

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

Computational pipeline

Multispectral extractor

- **Description**: Raw sensor output is converted to PNG and GeoTIFF format
- **Output**: `/ua-mac/Level_1/flirIrCamera`

See also

- Geospatial information

Multispectral imaging data

Summary

Multispectral data is collected using the Skye NDVI and PRI sensors. The normalized difference vegetation index (NDVI) is a simple graphical indicator that can be used to analyze remote sensing measurements.

Raw data access

Data is available via Clowder and Globus:

- **Clowder**:
  - ndviSensor collection
  - priSensor collection
- **Globus**:
  - `/sites/ua-mac/raw_data/ndviSensor`
  - `/sites/ua-mac/raw_data/priSensor`
- **Sensor information**:
  - Skye PRI collection
  - Skye NDVI collection

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

Computational pipeline

Multispectral extractor

- **Description:** NDVI binary files are converted to png thumbnail + geoTIFF\/netCDF
- **Output**: `/sites/ua-mac/Level_1/priSensor`

See also

- Geospatial information

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
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.

Phenotype data

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

Publication: http:\/\/dx.doi.org\/10.1016\/j.molp.2015.06.005

Demo Jupyter Notebook: https:\/\/github.com\/terraref\/computing-pipeline\/blob\/master\/demos\/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.

Data access

- **Clowder**: ddpscIndoorSuite
- **Globus**: `/sites/danforth/raw_data/<experiment name>`
- **BETYdb**: https:\/\/terraref.ncsa.illinois.edu\/bety

For details about accessing BETYdb, please see Data Access section.

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
    - 4x NIR side-view = NIR_SV_0, NIR_SV_90, NIR_SV_180, NIR_SV_270
    - 1x NIR top-view = NIR_TV
    - 4x VIS side-view = VIS_SV_0, VIS_SV_90, VIS_SV_180, VIS_SV_270
    - 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 (10 images) are inserted into a CSV file and added to the Clowder dataset
- If configured, the CSV will also be sent to BETYdb

3D Point Cloud Data

Summary

3D point cloud data is collected using the Fraunhofer 3D laserscanner. .

Data access

Data is available via Clowder and Globus.

- **Clowder**: scanner3DTop collection
- **Globus path**: `/sites/ua_mac/raw_data/scanner3DTop`
- **Sensor information**: Fraunhofer 3D scanner collection

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

Computational pipeline

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

ply2las extractor

- **Description**: PLY data is converted to LAS using the 3D point cloud extractor
- **Output**:
  - **Clowder:** LAS file is added to the dataset
  - **Globus**: `/sites/ua_mac/Level_1/scanner3DTop`

See also

- Geospatial information

How to access data

Overview

TERRA-REF data is available through four different approaches: Globus Connect, Clowder, BETYdb, 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.

For more information, see the Architecture Documentation.

Clowder

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

Raw data is transferred to the primary TERRA-REF compute pipeline on the Resource Open Geospatial Education and Research (ROGER) system using Globus Online. 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.

For more information, see Using Globus.

BETYdb

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

Use BETYdb to access about derived trait data.

For more information, see Using BETYdb.

CoGe

CoGe contains genomic information and sequence data.

For more information, see Using CoGe.

Other Data

- Field protocols
- Calibration protocols
- Field scanner operational log https:\/\/github.com\/terraref\/computing-pipeline\/issues\/128

Using Clowder

About Clowder

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.

See the Clowder documentation for more information about the software and its applications.

Requesting Access

To create an account, sign up at the TERRA-REF Clowder site and wait for your account to be approved. Once access is granted, you can explore collections and datasets.

Data organization

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.

Searching the database

Clowder allows users to search metadata and filter datasets and files with particular attributes. Simply enter your search terms in the search box.

Analyzing data in Clowder

Clowder includes support for launching integrated analysis environments from your browser, including RStudio and Jupyter Notebooks.

After selecting a dataset, under the "**Analysis Environment Instances**", select the "**Launch new instance with dataset**" drop-down, select the desired tool, then the "**Launch**" button. Select the "**Environment manager**" link to view the list of active instances. Find your instance and select the title link. This will display the tool with the selected dataset mounted. If you have a running instance, you can also "**Upload dataset to existing instance**".

Clowder Extractors

Through it's extractor architecture, Clowder supports automated computational workflows. For more information about developing Clowder extractors, see the Extractor Development documentation

Using Globus

About 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.

Installing Globus

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.

Requesting Access

To request access to the Terraref endpoint, send your Globus id (or University email) to David LeBauer (dlebauer@illinois.edu) with 'TERRAREF Globus Access Request' in the subject. You will be notified once you have been granted access.

Accessing Data via Globus

To transfer data to your computer or server:

1. Log into Globus https://www.globus.org
2. Add an endpoint for the destination (e.g. your local computer) https://www.globus.org/app/endpoints/create-gcp
3. Go to the 'transfer files' page: https://www.globus.org/app/transfer
4. 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/)

1. Click 'go'

2. Files will be transfered to your computer

See also

- Globus Getting Started
- Transfer API Documentation

Using BETYdb

About BETYdb

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).

Requesting access

To request access to BETYdb, register on the BETYdb web site. You will be notified once you have been granted access.

Data organization

The primary BETYdb Data Access Guide is largely relevant here, noting the following usages:

- Genotypes are stored in the `cultivars` table
- Plots are stored in the `sites` table. Plots are nested hierarchically based on geolocation.

Using the Advanced Search box

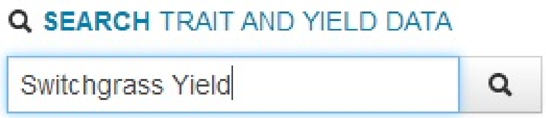
Most tables in BETYdb have search boxes. We describe below how to use the *Advanced Search* box to query data from these tables and download the results as a CSV file.

The Advanced Search box is the easiest way to download summary datasets designed to have enough information (location, time, species, citations) to be useful for a wide range of use cases.

(For more information about querying data from specific tables, see the BETYdb Data Access Guide.)

Using the Search Box

On the Welcome page of BETYdb there is a search option for trait and yield data (Figure 1). This tool allows users to search the entire collection of trait and yield data for specific sites, citations, species, and traits.



SEARCH TRAIT AND YIELD DATA

Switchgrass Yield

Figure 1: BETYdb Advanced Search Box

The *results* page provides a map interface and the option to download a file containing search results. The downloaded file is in CSV format. This file provides meta-data and provenance information, including the SQL query used to extract the data, the date and time the query was made, the citation source of each result row, and a citation for BETYdb itself.

Instructions

Using the search box to search trait and yield data is very simple: Type the site (city or site name), species (scientific or common name), cultivar, citation (author and/or year), or trait (variable name or description) into the search box and the results will show contents of BETYdb that match the search. The number of records per page can be changed to accord with the viewer's preference and the search results can be downloaded in the Excel-compatible CSV format.

The *search map* may be used in conjunction with search terms to restrict search results to a particular geographical area—or even a specific site—by clicking on a map. Clicking on a particular site will restrict results to that site. Clicking in the vicinity of a group of sites but not on a particular site will restrict the search to the region around the point clicked. Alternatively, if a search using search terms is done without clicking on the map, all sites associated with the returned results are highlighted on the map. Then, to zero in on results for a particular geographic area, click on or near highlighted locations on the map.

See also

- BETYdb Data Access Guide

Produced with Gitbook version 3.2.2

Using CoGe

CoGe contains genomic data.

About 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.

Requesting Access

TERRA REF Data and Software 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. Each type of data will include or point to the appropriate attribution policy.

Timing and Control of Release

We plan to release the data in stages or tiers. For pre-release access please complete the alpha tester application.

1. The **first tier** will be an internal release to the TERRA-REF team and the standards committee. This first tier release will be to initially quality check and calibrate the data and will take place as data sets are produced and compiled.
   i. By November 2016, it is an objective of the TERRA-REF team to establish a data release pipeline, wherein the release of data to this first tier will be within 21 days from the date of collection.
   ii. Access to the data will be arranged for by the resource producer (i.e. limiting access to selected users).
2. The **second tier** will enable the release of the data generated solely by the TERRA-REF team to other TERRA teams as well as non-TERRA entities.
   i. By November 2017, it is an objective of the TERRA-REF team to establish a data release pipeline, wherein the release of data to this second tier will be within 10 days from the data of collection.
   ii. It is noted that release of the data to the second tier may occur prior to publication and that access 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.
3. Access to the data will be determined by the resource producers and may be governed by separate license or other agreements. 1. iii)It is an objective of the TERRA-REF

team to enable the release of the data to the public by November 2018 but no later than the date of close-out of the awarded funds.

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 MIT or MIT compatible license (e.g. BSD, UIUC/NCSA, Apache v2).

Images, Phenotypes, and Other Raw Data

For other raw data, such as phenotypic data and associated metadata, we intend to release via Creative Commons with Attribution (CC by 4.0).

Contacts

Todd Mockler, Project/Genomics Lead (email: tmockler AT danforthcenter DOT org)

David LeBauer, Computing Pipeline Lead (email: dlebauer AT illinois DOT edu)

Erica Fishel, Technology Transfer Lead (email: efischel AT danforthcenter DOT org)

Nadia Shakoor, Associate Project Director (email: nshakoor AT danforthcenter DOT org)

TERRA REF 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 *(link?)*. 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 will make data available early as 'pre-release' with restrictive use policies. In these cases, people who wish to use the data for publication prior to official open release 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 TERRA REF website** (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 listserv (*to be determined @ terraref.org*).

   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:

   i. 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.'

   ii. 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. **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.

7. **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 NutNet paper, co-authors should recognize the final decisions belong to the lead author.

8. **Final manuscripts should be reviewed and approved by each co-author before submission.**

9. **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. An easy way to manage the author table is with the TERRA REF authorship template https://goo.gl/Z7qv4L.

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.

1. **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. Has each author made contributions in at least two areas in the authorship rubric? Did each author provide thoughtful, detailed feedback on the manuscript? Authors are encouraged to contact the TERRA REF PIs (Mockler, …) or authorship committee (Joe, Jane) about any confusion or conflicts.

Contributions Warrinting Co-authorship

Authorship must be earned through a *substantial contribution.* Traditionally, project initiation and framing, data analysis and interpretation, and manuscript preparation are all authorship-worthy contributions, and remain so for NutNet manuscripts. However, NutNet 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 coauthorship on opt-out papers, each individual must be able to check **at least two** boxes in the rubric, including contribution to the writing process. *These guidelines apply equally to manuscripts led by graduate students.*

Manuscripts published by TERRA REF will be accompanied by a supplemental table indicating authorship contributions. You can create and share a standard authorship table using google docs (https://goo.gl/Z7qv4L). For opt-in papers, a co-author is expected to have at least two of the following areas checked in the authorship rubric.

rubric item	example contribution meriting a checked box
Developed and framed research question(s)	Originated idea for current analysis of TERRA REF data; contributed significantly to framing the ideas in this analysis at early stage of manuscript
Analyzed data	Generated models (conceptual, statistical and/or mathematical), figures, tables, maps, etc.; contributed key components to the computing pipeline.
Contributed Data	generated a dataset being used in this manuscript's analysis.
Contributed to data analyses	Provided comments, suggestions, and code for data analysis
Wrote the paper	Wrote the majority of at least one of the sections of the paper
Contributed to paper writing	Provided suggestions such as restructuring ideas, text and citations linking to new literature areas, copy editing
Site level coordinator	Coordinated data collection, proofing, and submission of unreleased data for at least one site used in this manuscript.

Publications committee

Current co-chairs: TBD.

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.

Release / reprocessing schedule

We will release the data in stages or tiers.

First Tier

The first tier will be an internal release to the TERRA-REF team and the standards committee. This first tier release will be to initially quality check and calibrate the data and will take place as data sets are produced and compiled.

- By November 2016, it is an objective of the TERRA-REF team to establish a data release pipeline, wherein the release of data to this first tier will be within 21 days from the date of collection.
- Access to the data will be arranged for by the resource producer (i.e. limiting access to selected users).

Second Tier

The second tier will enable the release of the data generated solely by the TERRA-REF team to other TERRA teams as well as non-TERRA entities.

- By November 2017, it is an objective of the TERRA-REF team to establish a data release pipeline, wherein the release of data to this second tier will be within 10 days from the data of collection.
- It is noted that release of the data to the second tier may occur prior to publication and that access 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, provided that this is done in a timely fashion. Resource users should appropriately cite the source of the data and acknowledge the resource produces. The publication 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 contributing significantly to the work.
- Access to the data will be determined by the resource producers and may be governed by separate license or other agreements.
- It is an objective of the TERRA-REF team to enable the release of the data to the public by November 2018 but no later than the date of close-out of the awarded funds.

Technical Documentation

This section includes the following:

- Data Standards
- Directory Structure
- Data Storage
- Data Backup
- Data Transfer
- Data Processing Pipeline
- Data Collection
- Data Product Creation
- Sensor Calibration
- Quality Assurance and Quality Control

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

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.

Project Timeline

- August 2015: Establish committee, form a data plan
- January 2016: v0 file standards
- January 2017: v1 file standards, sample data sets
- January 2018: mock data cube generator, standardized data products, simulated data

- January 2019: standardized data products, simulated data

Data Standards Participants

- TERRA Project Representatives (6)
- ARPA-E Program Representatives (2)
- Board of External Advisors (6)

(numbers in parentheses are targets, for which we have funding)

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

Directory Structure

The data processing pipeline transmits data from origination sites to a controlled directory structure on the ROGER CyberGIS supercomputer.

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 eventually `Level_2` , etc) subdirectories.

When possible, sensor directories will be divided into days and then into individual datasets.

This directory structure is visible when accessing data via the Globus interface.

Data Storage

- Active Data Service: Active storage available for computing on campus cluster (or transfer to other HPC systems)
- ICEHouse: NCSA 10PB Tape Drive
- U of I Box (easy to use, 500GB)
- Roger: CyberGIS R&D server for GIS applications, 7PB storage + variety of nodes, including large memory. roger.ncsa.illinois.edu (CyberGIS; total 7 PB)

https://github.com/terraref/computing-pipeline/issues/87

Data Backups

Raw data

Running nightly on ROGER.

Script is hosted at: /gpfs/smallblockFS/home/malone12/terra_backup

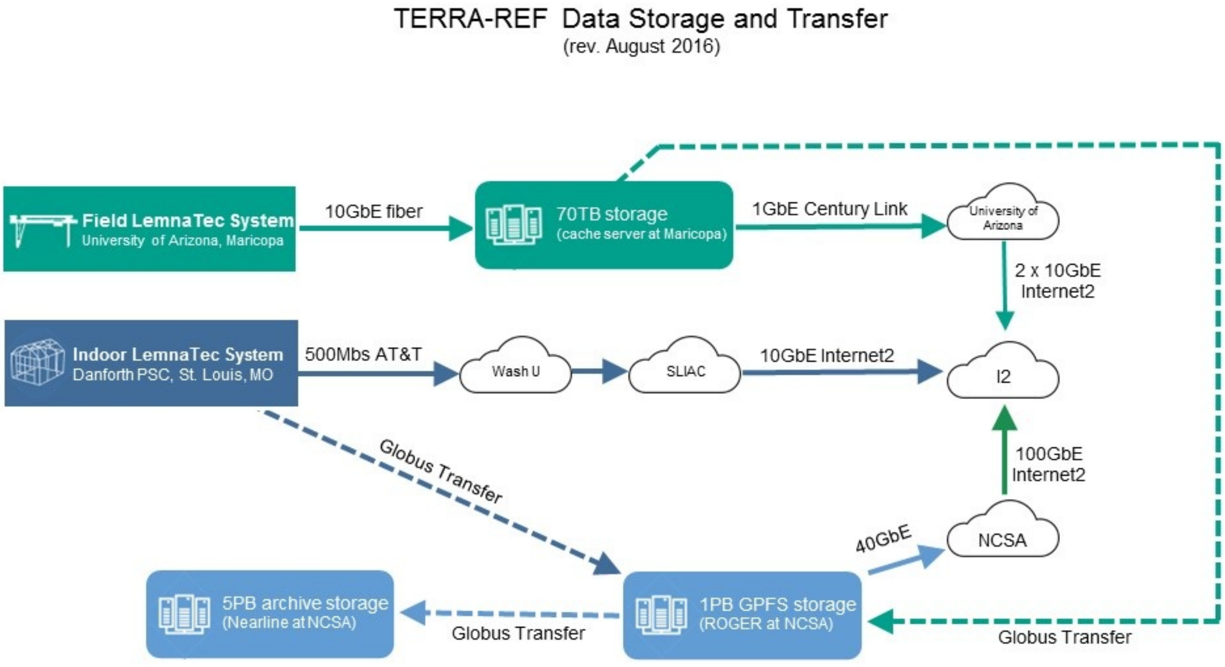
Script uses the Spectrum Scale policy engine to find all files that were modified the day prior, and passes that list to a job in the batch system. The job bundles the files into a .tar file, then uses pigz to compress it in parallel across 18 threads. Since this script is run as a job in the batch system, 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.



#Data Transfer

Maricopa Agricultural Center, Arizona

Environmental Sensors https:\/\/github.com\/terraref\/reference-data\/issues\/26

Using Logstash https:\/\/github.com\/terraref\/computing-pipeline\/issues\/106

Transferring images

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).

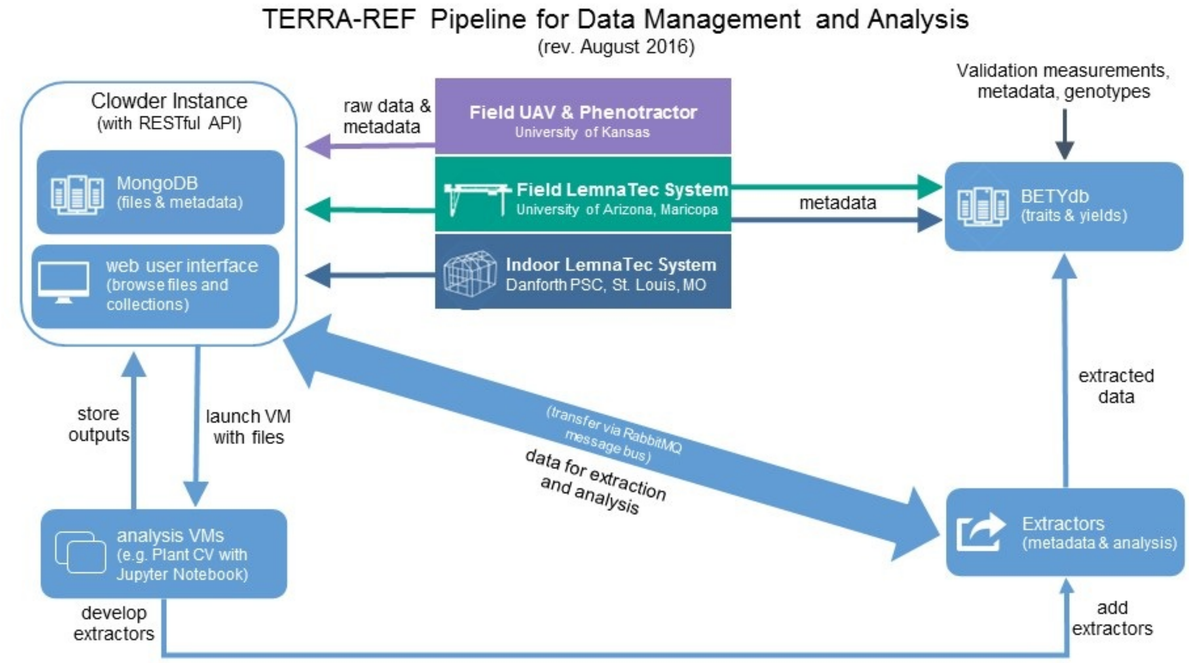
Transfer volumes

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

Kansas State University

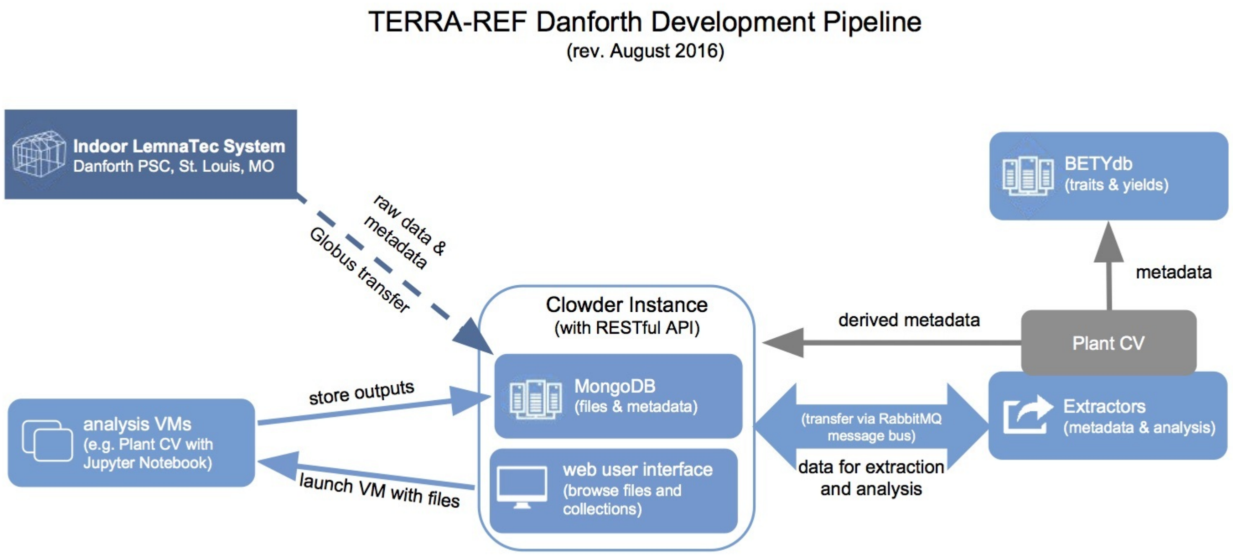
HudsonAlpha - Genomics

Data processing pipeline


TERRA-REF Pipeline for Data Management and Analysis
(rev. August 2016)

Maricopa Agricultural Center, Arizona

Automated controlled-environment phenotyping, Missouri


TERRA-REF Danforth Development Pipeline
(rev. August 2016)

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

How data was collected from each source

Maricopa Agricultural Center, Arizona

- The Lemnatec Scanalyzer Field Gantry System
  - Sensor missions
  - Scientific Motivation
  - What sensors, how often etc.
- Tractor
- UAV
- Manually Collected Field Data

https://docs.google.com/document/d/1iP8b97kmOyPmETQI_aWbgV_1V6QiKYLblq1jIqXLJ84/edit#heading=h.3w6iuawxkjl6 https://github.com/terraref/reference-data/issues/45

Automated controlled-environment phenotyping, Missouri

The Scanalyzer 3D platform 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.

Capturing images

LemnaTec Video Screnshot

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.

You can read more about the Danforth Plant Sciences Center Bellwether Foundation Phenotyping Facility on the DDPSC website.

Kansas State University

HudsonAlpha - Genomics

How data products were created

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[1] and NEON[2]

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

Downwelling Spectral Radiance Data

https://github.com/terraref/reference-data/issues/30

Genome annotations

Genome annotations are in GFF format GFF format contains genes and other genomic features. Allows “track” info for visualization http:\/\/useast.ensembl.org\/info\/website\/upload\/gff.html

Visualizing and annotating Genomes

Gbrowse is a comprehensive database + interactive web application for manipulating and displaying annotations on genomes.

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:

- Trimmomatic v 0.35
- bwa v 0.7.12-r1039
- samtools v 1.3.1
- picard-tools-2.0.1
- GATK v3.5-0-g36282e4

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=Sbicolor_313_v3.0.dict
```

Quality trimming and filtering of paired end reads

```
java –jar Trimmomatic-0.35/trimmomatic-0.35.jar PE -phred33 -trimlog trimlogPE.txt
SampleA_R1.fastq.gz SampleA_R2.fastq.gz SampleA_R1.PE.fastq.gz SampleA_R1.SE.fastq.gz
SampleA_R2.PE.fastq.gz SampleA_R2.SE.fastq.gz ILLUMINACLIP:adapters.fa:2:30:10
SLIDINGWINDOW:5:20 LEADING:20 TRAILING:20 MINLEN:60 2> trim.out
```

Aligning reads to the reference

```
bwa mem –M –R
“@RG\tIDSAMPLEA_RG1\tPL:illumina\tPU:FLOWCELL_BARCODE.LANE.SAMPLE_BARCODE_RG_UNIT\tLB:libra
ryprep-lib1\tSM:SAMPLEA” Sbicolor_313_v3.0.fa SampleA_R1.PE.fastq.gz SampleA_R2.PE.fastq.gz
> SAMPLEA.bwa.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_MAP=1000
REMOVE_DUPLICATES=true INPUT=SAMPLEA.bwa.sorted.bam OUTPUT=SAMPLEA.dedup.bam
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.intervals
```

Realign

```
java –Xmx8g –jar GenomeAnalysisTK.jar –T IndelRealigner –R Sbicolor_313_v3.0.fa –I
SAMPLEA.dedup.bam –targetIntervals SAMPLEA.realignment.intervals –o
SAMPLEA.dedup.realigned.bam
```

Variant Calling with GATK HaplotypeCaller

```
java –Xmx8g –jar GenomeAnalysisTK.jar –T HaplotypeCaller –R Sbicolor_313_v3.0.fa –I
SAMPLEA.dedup.realigned.bam --emitRefConfidence GVCF --pcr_indel_model NONE -o
SAMPLEA.output.raw.snps.indels.g.vcf
```

Hyperspectral Data Pipeline Overview

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)*. 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
```

QA/QC

Logging

Automated checks

visualizations

testing and continuous integration framework

https://github.com/terraref/computing-pipeline/issues/76

checking that scans align with plots https://github.com/terraref/computing-pipeline/issues/153

Development

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[1]

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:

   i. 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.

   ii. 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.

   iii. There is an `SE` heading without an `n` heading or vice versa.

   iv. There is neither a `yield` heading nor a heading corresponding to a recognized trait variable.

   v. 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.

   vi. 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.

i. 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.

ii. 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.

1.

> [1]. For additional documentation of the bulk upload feature, see the BETYdb data entry documentation. For information about submitting data via the API (not covered here), see Adding Traits via the Beta API. ↩

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/icommands.x86_64.tar.bz2
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

Extractors are services that run silently alongside Clowder. They can be configured to wait for specific file types to be uploaded into Clowder, and automatically execute operations on those files to extract metadata.

It is possible to develop extractors for new file types or tasks. See the NCSA Extractor Development wiki for complete instructions

Setting up a test environment for development

The NCSA Clowder wiki provides an up-to-date installation guide for Clowder.

In a fresh installation, Clowder is not configured with an email server - it does not send confirmation when someone registers for an account the confirmation email will not be sent correctly. However the Clowder console will still display the contents of the email, so the confirmation URL can be copied from there.

References

- **Source code** is available as a collection of Git repositories.
- Tutorial
  - overview
  - video
- pyClowder is designed for this purpose.
- Development in Windows
- Using Clowder via National Data Service interface

- Contacts: Max Burnette via email, phone, on GitHub, or on our Slack Channel

Moved to https://github.com/terraref/tutorials

Contributor 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/

Web-based collaboration tools

For use by the TERRA Reference Phenotyping Standards Committee.

Overview

All of the web-based software below provides the ability to organize projects hierarchically, facilitate sharing, and support collaboration. Much of this is publicly viewable.

Core Communication Tools

- **Github** github.com/terraref project management, website content and hosting, collaborative software development
- **Google Drive** collaborative editing of documents that we create (notes, manuscripts, etc)
- **Slack:** terra-ref.slack.com (signup)

GitHub

TERRA Ref sites on Github:

- **Data products repository** https://github.com/terraref/reference-data
  - issues and milestones: https://github.com/terraref/reference-data/issues
- **Computational Pipeline Repository** https://github.com/terraref/computational-pipeline
  - issues and milestones: https://github.com/terraref/computational-pipeline/issues
- **Website for R&D** : https://terraref.ncsa.illinois.edu
- **Documentation**
  - GitHub Repository: https://terraref.ncsa.illinois.edu
  - Edit in the GitBook Desktop Editor or GitBook Web interface (see GitBook Documentation)

Using Github:

- **Features**

  - Interface to 'git', a specialized command-line tool for version control.
  - code management / collaboration:

  - Issue tracking and discussion forum https://guides.github.com/features/issues/
    - participants can reply to issues via email, similar to an email discussion list

- **GitHub Documentation**

- **GitHub Desktop**

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