Get Grounded in TERRA-REF: Publicly-accessible, Hi-res Sensor Data for Crop Phenomics

David LeBauer, PhD
Director of Data Science, Arizona Experiment Station
@dlebauer
Overview

• High Throughput Phenomics, Briefly
• Why A Reference Dataset?
• The Field Scanner
• Sensor and Phenotype Data
• Accessing the data
• Toward Open Phenomics Pipelines
Ag and Open Research: A Long and Successful History

Norman Bourlaug

Land Grant Universities

CGIAR Research Centers

THE UNIVERSITY OF ARIZONA
College of Agriculture & Life Sciences
High Throughput Phenotyping

• Phenotypes:
  • Plant traits; anything you can measure.

• High Throughput Phenotyping:
  • *Increased frequency and resolution.*
  • *Greater variety* of measurements.

• Challenges:
  • Sensors are expensive.
  • Data are difficult to interpret.

• ARPA-E TERRA program major investment to advance HTP.
ARPA-E TERRA Program Vision

**CATEGORY 5: PROGRAMMATIC REFERENCE DATA GENERATION AND HOSTING**

The successful applicant for Category 5 will generate and host for TERRA teams reference phenotyping data through an energy sorghum field test plot monitored by a state-of-the-art field phenotyping system provided by ARPA-E.

...sensors. At this time, it is the intent of ARPA-E for the Category 5 awardee to release all data to the public at an appropriate time determined by ARPA-E that is consistent with DOE regulations.

**FUNDING OPPORTUNITY ANNOUNCEMENT**
TERRA REF Data and Computing Objectives

- Reference sensor, phenotype, and genomics data
- Open source pipeline
- Open access reference data
- Open computing
Stereo

Hyperspectral

FLIR

3D Laser

PS 2
Field Scanner Data

- High Resolution: daily to weekly @ mm²
- TB per day from sensors
- Hundreds of traits, millions of observations per season
- Genomics: 40x coverage of 400 genotypes
- Open (available now; public release spring 2020)

Image credits: N. Fahlgren, C. Zender, S. Hajmohammadi, Z. Li, R. Pless, D. LeBauer, TERRA REF team
High resolution RGB
Full Field Mosaic
Thermal Infrared
Laser 3D

g.png

l.las

p.ply
Visible / NIR Hyperspectral

% Reflectance at 530 nm
PSII Camera

F0

Fv

Fm

Fv/Fm

https://github.com/nfahlgren/shared_jupyter_notebooks/blob/master/terraref/psII_analysis_demo.ipynb
Soil Maps

12 size classes sand, silt, clay hydraulic properties
Moisture content
Organic matter
pH

Physicochemical and Hydrologic Characterization TERRA-REF
South Field (in prep)
Ebrahim Babaeian, Juan R. Gonzalez Cena, Mohammad Gohardoust, Xiaobo Hou,
Scott A. White, and Markus Tuller
Other Data

• Agronomic Metadata:
  • Management, Plot Boundaries, Experimental Design

• Soil Maps

• Weather
  • Time series of temperature, precip, irrigation...

• Hand Measurements

• Derived traits

• Genomics:
  • VCFs & hapmaps
Plant mask / canopy cover
Canopy cover extractor

L0: Raw image
b829-019c457efb50_left.bin
b829-019c457efb50_metadata.json

L1: Georeferenced

L2: Plant Mask

L3: Plot Level Canopy Cover

3 Channel Threshold
Morphological Enhancement
Adversarial (ML) Enhancement

THE UNIVERSITY OF ARIZONA
College of Agriculture & Life Sciences
Uses of TERRA REF Data

• Crop Improvement
• Sensor Evaluation and Fusion
• Basic Biology: Phenotype = Gene x Environment
• Teaching
• Basic ML Research
TERRA REF for ML Research

DARPA D3M

NSF IDEAS Lab

ARPA-E OPEN
Our goal is to make this as useful as possible

Who is it available to?
- Anyone
- Published Data = Public Domain
- Other Data = Data Use & Authorship policy

Publication
- Data Publication *in prep (Spring 2020)*
- Initial release: Seasons 4 and 6
- Other data available as *needed*
Lots of ways to Access Data

<table>
<thead>
<tr>
<th>Resource</th>
<th>GUI</th>
<th>API*</th>
<th>clients</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensor Data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Globus</td>
<td>Yes</td>
<td>Yes</td>
<td>R, Python</td>
</tr>
<tr>
<td>Clowder</td>
<td>Yes</td>
<td>Yes</td>
<td>Python</td>
</tr>
<tr>
<td><strong>Trait Data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BETYdb</td>
<td>Yes</td>
<td>Yes</td>
<td>R, Python</td>
</tr>
<tr>
<td><strong>Genomics Data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CyVerse</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>CoGe</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

*GUI: Graphic User Interface
API: Application Programming Interface

docs.terraref.org/user-manual/how-to-access-data
Browsing Available Phenotypes: Traitvis

TERRA-REF Trait Data

MAC Season 4  MAC Season 6  MAC Season 9

All BAP With Late Season Drought  Border Plots

Variable
Canopy Cover (%)  Cultivar
None

Map  Download

Canopy Cover (%)

Date
2017-04-13  2017-09-21

How to access data
You can access Canopy Cover (%) data for MAC Season 4, using either:
1. BETYdb API
2. R traits package

API key
Both methods will require an API key.
You can get an API key by signing up for the TERRA Ref BETYdb traits database. Register for BETYdb at
https://terraref.ncsa.illinois.edu/betydb/

R traits package
Install the traits package from CRAN using: install.packages("traits").
You can then use the following chunk of R code to access the data:

library(traitis)

options(betydb_url = "https://terraref.ncsa.illinois.edu/betydb", betydb_api_version = "beta", betydb_key = "YOUR_API_KEY")
canopy_cover <- betydb_query(trait = "canopy_cover", sitename = "Season 4", limit = "none")

BETYdb API
You can also access the data at this URL:
https://terraref.ncsa.illinois.edu/betydb/api/search?trait=canopy_cover&sitename=Season 4&limit=name&key=YOUR_API_KEY.
See API documentation for details.
For browsing files accessing metadata: Clowder
Tutorials: Quick Start Access Data

TERRA REF Tutorials

David LeBauer and others

2020-01-31

Chapter 1  Overview

This book is intended to quickly introduce users to TERRA REF data through a series of tutorials. TERRA REF has many types of data, and most can be accessed in multiple ways. Although this makes it more complicated to learn (and teach!), the objective is to provide users with the flexibility to access data in the most useful way.
Tutorials: Webinars and Videos

Traits 4: How to get TERRA REF personal API key
Traits 3: How to plot TERRA REF trait data
Traits 2: How to select TERRA REF trait data by trait
Traits 1: How to download TERRA REF trait data
CyVerse VICE: Visual Interactive Computing Environment
CyVerse VICE:
Rstudio and Jupyter notebooks
For Transferring Large Files: globus.org #TERAREF
Next Steps: Migrating to CyVerse, Distributed Computing

Chris Schaufer, Applied Cyberinfrastructure Concepts (Nirav Merchant & Eric Lyons’ Class)
github.com/uacic/PhytoOracle
phytooracle.readthedocs.io/en/latest/abstract.html
Algorithm Templates

Enable Best Practices
- Standard formats and vocabularies
- Versioning
- Publication
- Reuse
- Testing
- Documentation

Minimize Overhead
- Abstract computing environment
- Standardize data / metadata handling

1. Setup: Click the Use this template clone
2. Definitions: Fill in and modify the definitions
3. Algorithm: Replace the code in the template
4. Test: Run the testing.py script to run tests
5. Generate: Run generate.py to create a Docker image for you
6. Docker: Create a Docker image for you
7. Finishing: Finish up your development
Algorithm Templates

```python
def calculate(pxarray: np.ndarray):
    r"""Calculates one or more values from plot-level RGB data
    Arguments:
    pxarray: Array of RGB data for a single plot
    Return:
    Returns one or more calculated values
    """
    # ALGORITHM: replace the following lines with your algorithm
    channel_size = pxarray[:, :, 1].size

    # RETURN: replace the following return with your calculated values.
    return channel_size
```

Algorithm Libraries

v1: github.com/terraref
v2: github.com/agPipeline

Kooper, Burnette (NCSA)
Schnaufer (UA)

Your Algorithm Here
Data Formats, Standards & Conventions

**Sensors**
- CF Conventions
- OGC
- geoTIFF
- NetCDF-CF
- LAS

**Traits**
- BETYdb
- Crop Ontology
- Agronomy Ontology
- AgMIP/ICASA/
- BRAPI

**Genotypes**
- Genesys
- Accession

**Genomics**
- BAM, FASTQ,
- VCF, BED,
- FASTA, GFF

**Selection**
Data and Software Interoperability: Agricultural Research Data Network (ARDN)

With Hoogenboom, Abendroth, Basso, Parr, Vellidis
Acknowledgements

Todd Mockler
Nadia Shakoor
Noah Fahlgren
Stephen Kresovich
Jeremy Schmutz
Jesse Poland
Geoff Morris
William Rooney
Pedro Andrade-Sanchez
Duke Pauli
Michael Ottman
Markus Tuller
Max Burnette
Rob Kooper
Jeff White
Robert Pless
Roman Garnett

Chris Schnaufer
Sateesh Peri
Michele Cosi
Ebrahim Babaiean
Eric Lyons
Nirav Merchant
Julian Pistorius
Emily Cain
Kristina Riemer
Emmanuel Gonzalez
Tony Edgin
Tyson Swetnam

and many others!
CyVerse is supported by the National Science Foundation under Grants No. DBI-0735191, DBI-1265383 and DBI-1743442.