2014 TO 2016 PHENOTYPIC DATA UPDATE

Naser B. AlKhalifah

Overview

1. Introduction
2. Data Curation Workflow
3. Inbred/Hybrid Phenotyping Data Problems and Solutions
4. Data Curation Progress and Current Efforts
5. Future Considerations and Projects
6. Acknowledgments
7. Questions

Introduction

Education:

- M.S. Plant Breeding (Agronomy) – Iowa State University
- B.S. Genetics (Liberal Arts and Sciences) – Iowa State University

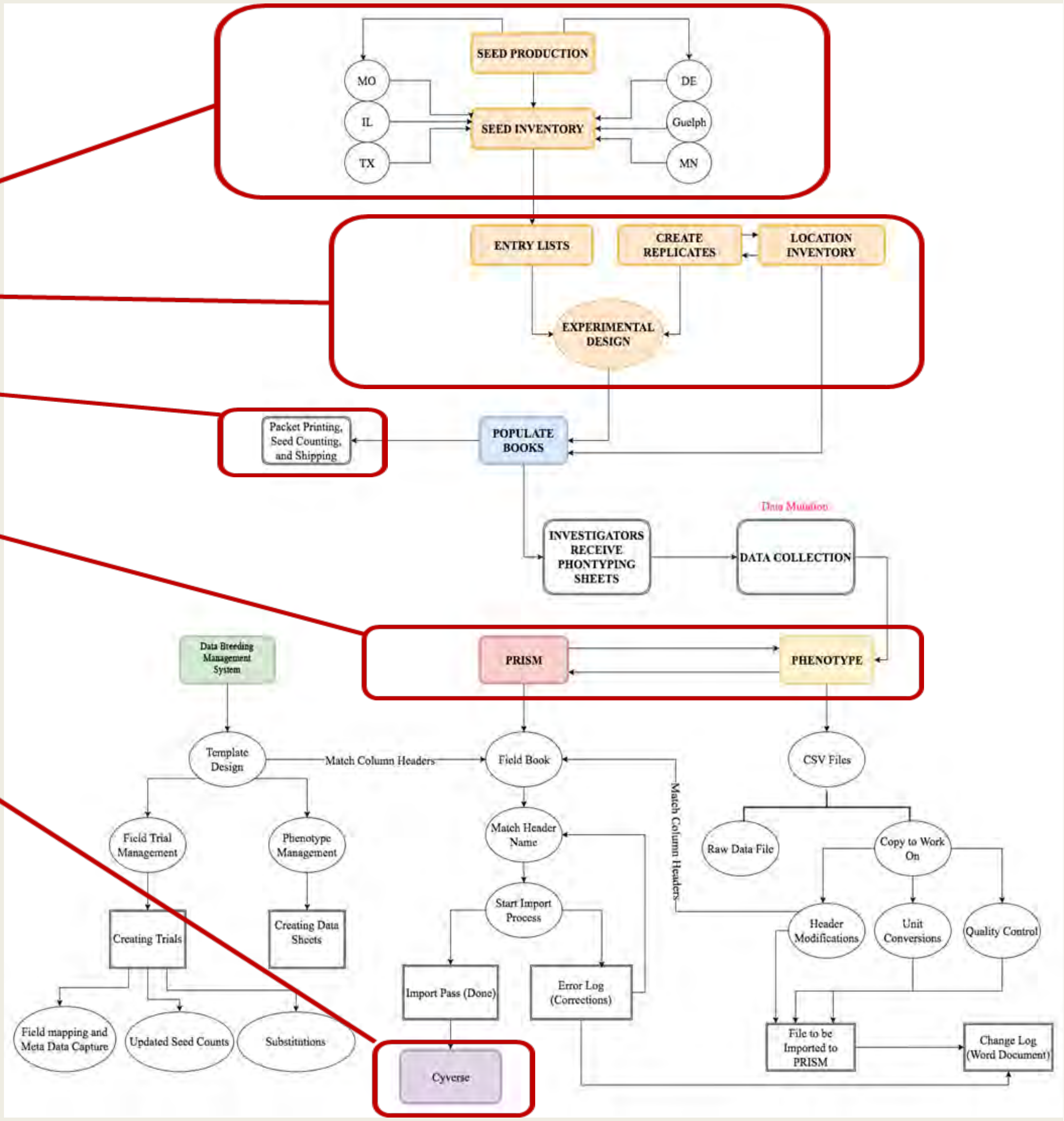
Professional Experience:

- Assistant Scientist – Iowa State University (current)
- Field Research Associate (II) – Monsanto Company (2014-2016)
- Lab Technician (II) – High-throughput Genotyping Lab: Monsanto Company (2009-2010)
- Seasonal Worker – Pioneer Hi-bred Intl. (2009)

Introduction Cont’d

Involvement:

- Planting locations and seed inventory
- Experiment planning & building
- Seed logistics
- Data curation
- Data release



Data Curation Workflow

Data Curation Workflow

1. Seed and location inventory received.
2. Fieldbooks are created within the PRISM software.
3. Seed packaging, shipping, and substitutions are made.
4. Updated fieldbooks are uploaded to both the Google Workbook and Discovery Environment (DE).
5. Updated fieldbooks & metadata files are returned through Google Sheets and DE.
6. Data are aggregated and curated and re-imported back into PRISM.
7. Metadata are curated and saved on a local server.

Inbred/Hybrid Phenotyping Data Challenges and Solutions

Challenges and Solutions Concerning Phenotypic Data

- Problem: Data fragmentation across multiple storage platforms.
  - Cyverse: Discovery Environment (DE)
  - Google Sheets
  - USDA Servers
  - Dropbox

- Solution: Consolidated all phenotypic data for years 2014 through 2016 into **one database** (currently using PRISM)

Challenges and Solutions Concerning Phenotypic Data Cont’d

- Problem: Critical plot identifiers missing from returned data file.
  - Removal of Rep and Plot or re-numbering of Rep and Plot columns.

- Solution: Created an alternate plot column to ensure that plot associations are maintained.

Challenges and Solutions Concerning Phenotypic Data Cont'd

- Problem: User changes to fieldbook column headers for commonly collected phenotypes across locations.
  - *Silking – Female Flowering (FF)*
  - *Anthesis – Pollen Shed (PS)*

- Solution: Created an import and export template (for each year) that automatically detects and matches column headers in the PRISM software.

Challenges and Solutions Concerning Phenotypic Data Cont'd

- Problem: Inconsistent height measurement scale across locations
  - Some locations report in metric while others report in imperial.
  - Placing ear height data in plant height column and vice versa.
  - Switching scales within the same column.

- Solution: Developed a program that detects when the imperial scale is used for plant and ear height measurement and converts to metric scale.

Data Curation Progress & Current Efforts

Progress

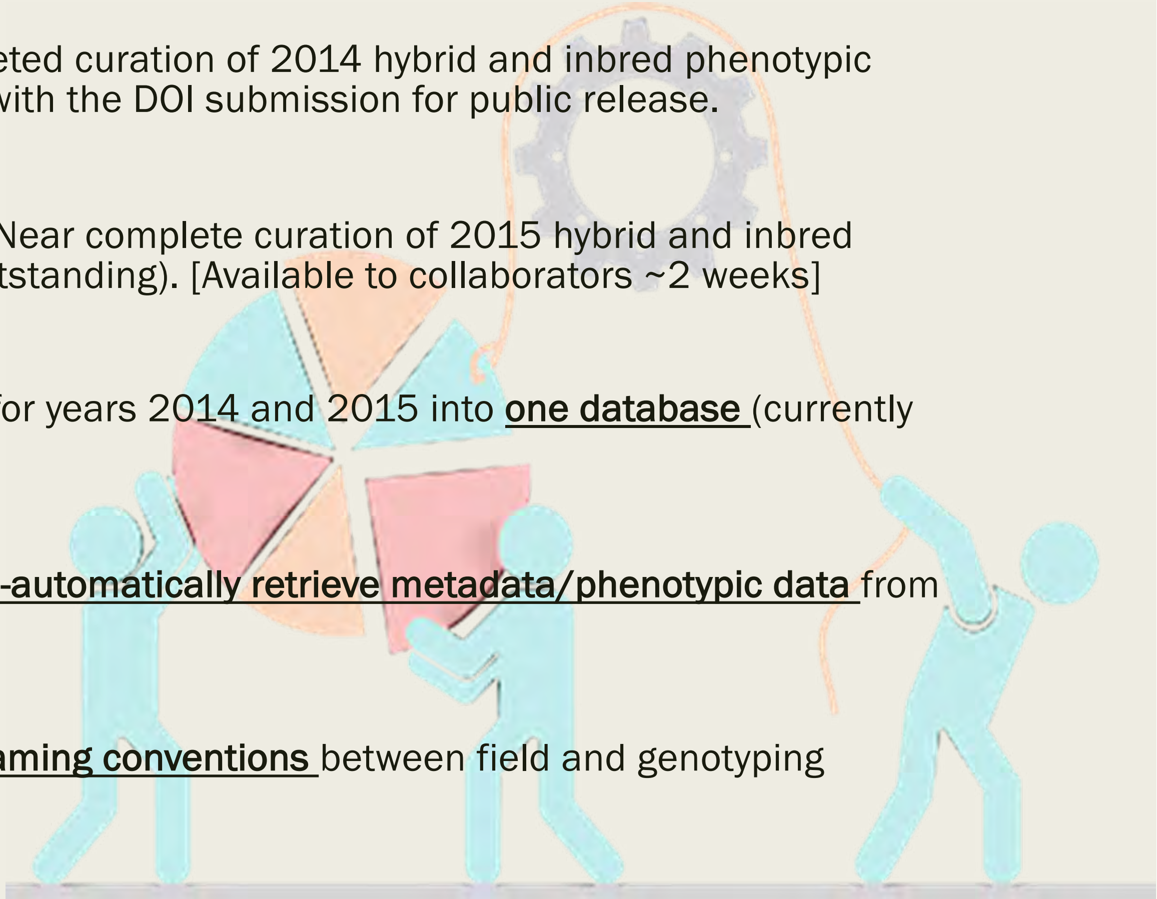
- 2014 public release (DOI): Completed curation of 2014 hybrid and inbred phenotypic and genotypic data and assisted with the DOI submission for public release.

- 2015 collaborator release (ARK): Near complete curation of 2015 hybrid and inbred phenotypic data (two locations outstanding). [Available to collaborators ~2 weeks]

- Consolidated all phenotypic data for years 2014 and 2015 into one database (currently using PRISM).

- Developed a SAS program to semi-automatically retrieve metadata/phenotypic data from Google Workbook.

- Created and applied consistent naming conventions between field and genotyping efforts.



Current Efforts

- Aggregate and curate 2016 phenotypic data and import into database.
- Create field books for 2017 locations and forward to collaborators.
- Assist with seed packaging efforts and account for substitutions.
- Develop a program that can automatically extract and reformat data for seamless integration with PRISM.

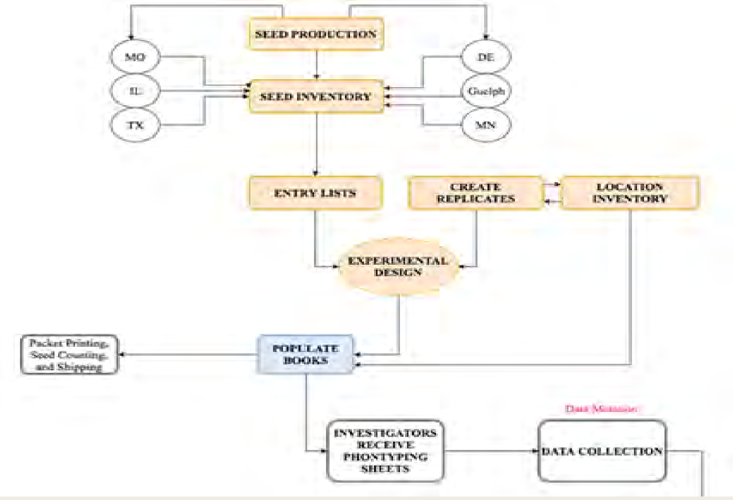
Data Release Timeline

Month	Year	Release to Collaborators	Public Release
February	2017	2016	2015
February	2018	2017	2016
February	2019	2018	2017
February	2020	2019	2018

Future Considerations and Projects

Future Consideration: Seed & Location Inventory

- Implement a system that ensures accurate and rapid transfer of location inventory.
- Implement a system that ensures accurate and rapid transfer of information between seed producers and data team.
- Set systems to track and implement substitutions during seed packaging.



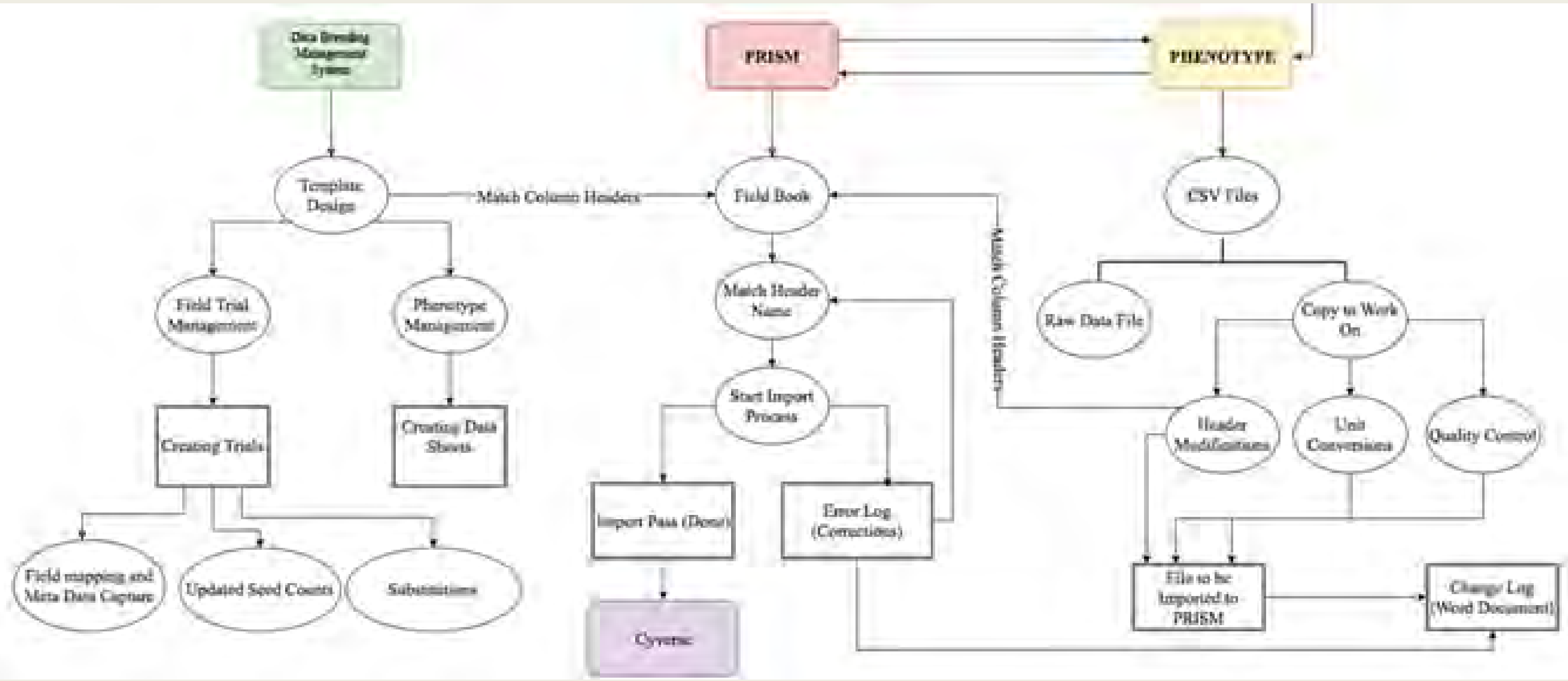
Future Projects: Statistical Analysis Software (SAS) for retrieving metadata from Google Workbook

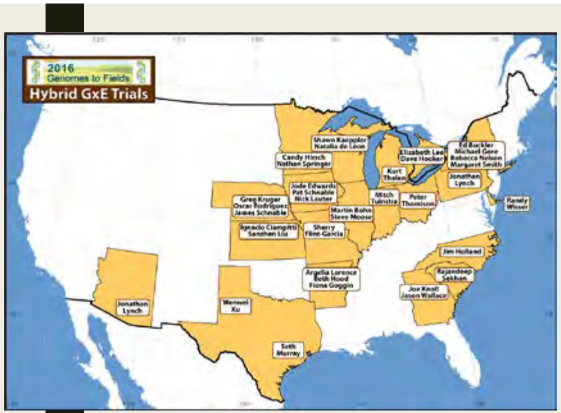
- Develop a macro to automatically retrieve metadata from Google Workbook as soon as data is available.

- Develop a macro which automatically transposes data to a format that is PRISM accessible.

- Use SAS as a means to track data changes instead of PRISM error reports.

Future considerations: Open-Source Database





Acknowledgments

G2F Executive Committee

- Pat Schnable (Iowa State Univ), co-lead
- Natalia de Leon (Univ of WI), co-lead
- Ed Buckler (USDA/Cornell)
- Shawn Kaeppler (Univ of WI)
- Jonathan Lynch (Penn State Univ)
- Nathan Springer (Univ of MN)
- David Ertl, Iowa Corn Growers' Association

IOWA CORN™

USDA ars

NCGA

IOWA STATE UNIVERSITY
Plant Sciences Institute

maizeGDB
MAIZE GENETICS AND GENOMICS DATABASE

Questions to You

- **EASY DATA COLLECTION** Do you prefer to use the Google Workbook or the Discovery Environment for collecting field data? What field collection devices and platforms do you currently use?

- **ADDITIONAL TRAITS** Are there any additional traits you would like to see us include within the fieldbooks?

- **EASY DATA ACCESS** What are your thoughts on the annual DOI, ARK for staging, etc.

Questions