IBDB Phenotype Data Management System

Data Modeling

Tucson

January 21-23, 2013

Table of Contents

I. Introduction

From Jan 21-23, 2013, a team met for an in-person, intensive working session to create a logical data model and being work on a physical schema for the new IBDB Phenotyping Database.

The team used user stories to define the boundaries of the modeling problem. The team began mapping the logical data model to the Chado/Chado ND physical schema and started developing the business rules to govern the mapping. This document contains the results of the proceedings.

In terms of design criteria, the new IBDB Phenotyping Database must be plant breeding centric, extensible, modular and high performance. Further, our ideal solution would be based on existing solutions that we can leverage and that are accepted and open. Prior to the meeting, we narrowed this down a list of available schemas to three likely candidates: Chado with ND, Katmandoo and Germinate. All three are available under open license and map closely to the plant breeding problem. See **Appendix D – Schema Candidates**.

II. User Stories

The first day of the meeting was dedicated to aligning expectations and establishing a baseline description of the problem space. The goal was to establish clearly the problem the phenotyping data model will solve.

The team used the four step breeding problem as the basis for scope analysis.

1. Logistics and Characterization data for Population Development
2. Evaluation Data for Advanced Lines
3. Derived and Analyzed Data
4. Cross-study Breeders' queries

Additionally the team considered the case where breeders want to build a general query. It was decided that while this modality is very important to certain breeders, its undefined scope made it too difficult to include in capturing User Stories. Some sample queries of this type should still be included in the performance testing of the final physical schema.

The team developed a list of user stories that will drive development of the logical and physical data model for a Phenotyping Data Management System.

After completing user needs analysis, the team devoted some time to desired extensibility of the data model, performance parameters, hardware baseline, and physical requirements of the database (e.g. local v. cloud hosting).

A. Logistics and Characterization Data for Population Development

In this step, it is assumed the breeder has already selected parental germplasm for development of breeding populations. This step focuses on data entry pertaining to the population development process (selected lines or families and methods of selection) as well as breeders notes and some basic observational phenotypic data.

The breeder may perform queries about the parents in the GMS and Pedigree system. These queries are out of scope for the Phenotyping database.

The following data entry user stories are relevant to this step in the breeding process.

A-1. As a breeder, I can enter study and location properties (e.g. soil PH) about my breeding project.

A-2. As a breeder, I can enter a list of crosses, with identifying information, in my breeding project.

A -3. As a breeder, I can record plants and rows to be harvested or discarded for the next generation.

A-4. As a breeder, I can record notes and basic phenotypic observations on lines or families in my breeding populations.

B. Evaluation Data for Advanced Lines

In this step in the process, the breeder is ready to begin designing field trials and associated Fieldbooks. This step focuses on data entry about the field trials (experiments) the breeder is developing and the measurement data from the trials. The breeder must also query data within a single field trial for analysis.

In addition to A-1 above, the following data entry and query user stories are relevant to this step in the breeding process.

B-1. As a breeder, I can enter experiment properties (e.g. design and treatment information) about my field trial.

B-2. As a breeder, I can record measurement data about germplasm in my field trial.

B- 3. As a breeder, I can look up a subset of data from my breeding project.

B-4. As a breeder, I can perform analysis on data from a single trial in my breeding project and store the results.

C. Derived and Analyzed Data

In the third step in the breeding process, the breeder performs local quality assurance and future analysis on data from the field trial(s). This step focuses on queries of data from a single site and entry of analysis results(e.g. mean) data.

In addition to those listed above, the following data entry and query user stories are relevant to this step in the breeding process.

C-1. As a breeder, I can query basic phenotypic data and derive values for new observations based on transformations of the basic observations.

C-2. As a breeder, I can perform analysis on data from individual trials in my breeding project and store the results.

C-3. As a breeder, I can perform analysis on data from multiple trials (environments) from the same study, possibly using means data from single trial analysis for the same study

D. Cross-study Breeder's Queries

In the final step in the breeding process, the breeder has completed the field trial and now wants to use information from multiple projects to choose new parents to bring into the breeding program as well as decide which lines to advance to the next stage of testing or release to farmers.. This step focuses on queries from multiple studies.

There is a single user story in this step with multiple queries associated. The team identified the core set of relevant queries and they are available in Appendix A – Cross-Study Queries.

D-1 As a breeder, I can perform analysis on data from multiple trials across multiple breeding projects.

Included for completeness, but not given tremendous weight in the data modeling:

D-2 As a breeder, I can perform free-form queries to retrieve data across multiple breeding projects. This involves selecting variables for which a report is required, and then defining a filter based on (possibly different) variables to specify the observation units from different studies for which the report is required.

III. Data Modeling

The second day of the meeting was devoted to developing a logical data model for the Phenotyping Data Management System. Using the user needs analysis from day one as problem boundaries, the team discussed the desired properties of the logical data model.

1. Flexible, to allow breeders to perform their work and record their data according to the needs of their crops and projects.
2. Extensible, to allow new crops and subdomains to be added seamlessly in the future.
3. Modular and based on a domain-centered model to support the wide array of tools development.
4. High Performance, supporting cross-study queries.

Prior to the meeting, the team had narrowed down the list of schema to three candidates most likely to fulfill the needs of the Pheontyping Data Management System: Chado with ND extensions, Katmandoo and Germinate.

The team elected to start modeling first using Chado with ND extensions, partly due to the attendance of a Chado expert. Katmandoo and Germinate are still under analysis.

Chado meets the desirable properties of the logical data model. It provides for flexibility through the use of ontologies and controlled vocabularies to extend the range of entity tables through the use of property tables. It is designed to be extensible and has already been extended via the Chado ND module among others. The core tables map to the Data Management domains of Germplasm, Study, Location, Ontology, Experiment and Phenotype. The design of the Experiment table is polymorphic, to allow additional flexibility in trial design. The modular design is expected to support reasonable performance, but this will need to be tested and optimized.

Because of its flexibility, any use of Chado in the IBDB would need to be coupled with a well-defined, rigidly enforced set of business rules. The business rules define the appropriate usage of the Chado tables, while still allowing for flexibility for breeders at the project, location and trial (experiment) design level.

The team proceeded by data modeling a small sample problem to flesh out the usage of Chado and associated business rules. The initial list of business rules is provided in Appendix B – Business Rules. The team was able to complete a large percentage of the core business rules during the meeting and will continue developing them using two additional sample problems. See Next Steps below.

IV. Next Steps

- Meeting Report – Feb 1st, 2013
- Implementation of Chado in MySQL in IBDB Schema – Jan 29th, 2013
- Cost and Schedule to migrate IBDB to Postgres (information purposes only at this time) – Date TBD
- Further define and finalize schema and business rules
    - Populate Chado with Breeding Problem 1 – Feb 2nd, 2013
    - Populate Chado with Breeding Problem 2 – Feb 9th, 2013
    - Populate Chado with Breeding Problem 3 – Feb 9th, 2013
- Complete Analysis of Remaining Schema
    - Germinate – Feb 8th, 2013
    - Katmandoo – Feb 8th, 2013

V. Appendix A – Cross-Study Queries

The cross-study queries can be organized into two logical groupings, simple and complex. The results of simple queries typically serve as input to complex queries and breeders may wish to examine and filter them prior to performing the complex queries. The delineation between simple and complex is somewhat arbitrary.

A. Simple Queries

1. Find all germplasm with a certain phenotypic characteristic.
2. Find available data for given list of germplasm.
3. Find values for a set of traits given a set of germplasm.
4. Find all populated traits for selected germplasm [at a given location].
5. Find all locations where a list of germplasms was together.
6. Find all locations where a trait was measured for a list of germplasms.

B. Complex Queries

1. Find germplasm where a given trait(s) had "suitable" performance for a target environment.
2. Compare performance of given pairs of lines, by trait. Inputs to the comparison are means, number of trials and summary statistics from trials where the pairs of lines were evaluated together.
3. For a given germplasm and set of locations, show the trait values for the germplasm per location and average trait values over all germplasms at that location.

VI. Appendix B – Business Rules

A. Table Mappings

This section describes the mappings from the logical data model to the Chado and Chado ND schema for the IBDB Phenotyping Database. This is the understanding of the rules as they stood at the end of the meeting on Jan 23, 2013. Eventually these rules will need to be captured formally as a data dictionary, ontology and defined usage.

The intent of the description is to be as readable as possible, so in some places the relationships are described at their logical level. For example, property type names may be stored in a separate table with links to the property table by id. For clarity a description of the property table would describe the type by its text name rather than id.

Primary keys are omitted in all descriptions of tables.

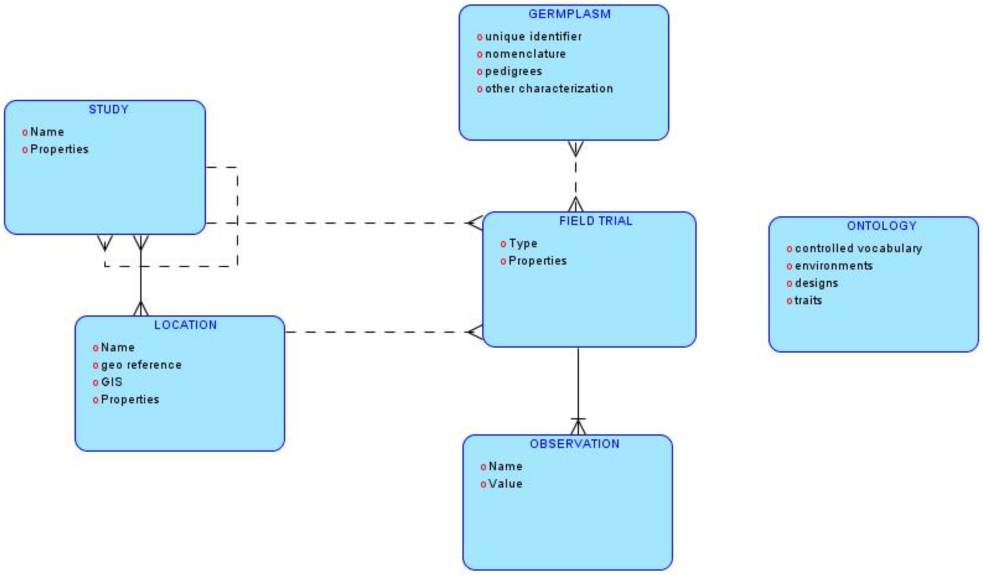

FIGURE I – SIMPLIFIED LOGICAL DATA MODEL

1. **Project Table (Study)**
   A Study is captured using the Project table. Information stored at this level describes properties relevant for all field trials in a Project (Study). Since it is important both that local breeders are free to use their nomenclature and that these local terms are mapped to a central ontology, the properties table maps all terms to the Ontology at the project level. Example property terms include name of the Study, Start/End Date, Principal Investigator, Phenotypic data types, etc.

   The project_properties table captures links to the ontology. Properties with the same rank represent a single compound property. The type column distinguishes the components of the compound property. Note that these entries are just the names of the properties with their ontology mappings. Values for properties are stored in the appropriate sub-module, usually in Experiment.

TABLE 1: EXAMPLE PROJECT PROPERTIES FOR PI AND GRAIN YIELD

Project_id	Type (would use an ID)	Value	Rank
1	Local Name	Principal Investigator Name	1
1	Link to Ontology table	2	1
1	Short name	PI	1
1	Local Name	Grain yield	26

1	Link to Ontology table	12	26
1	Short Name	YIELD	26

2. **Geolocation table (Location)**
The Geolocation table maps to the Location module of the logical data model. Information in this table corresponds to actual physical locations where Field Trials are conducted. Information about Locations is stored in the GMS database, outside the scope of the Phenotyping database. The Geolocation table has a link to the GMS location table in the existing schema via a property linked to the loc_id.

   Locations are linked directly both to Project (study) and Experiment (Field Trial).

3. **Stock table (Germplasm)**
The Stock table maps to the Germplasm module of the logical data model. Entries in the Stock table represent the Germplasm used in a field trail. Information about the Germplasm is stored in the GMS database, outside the scope of the Phenotyping database. Thus, the Stock table serves as a link to the information in the GMS.

   The Stock table is linked to the Germplasm table in GMS. The uniquename column contains the IBDB_GID. This column uniquely identifies a germplasm, but is not necessarily unique in the Stock table. Type = entry and the location-based name for the germplasm (e.g. entry_1) is stored in the name column.

4. **Experiment table (Field Trial)**
The Experiment table maps to the Field Trial Module in the Logical Data Model. Experiments (Field Trials) and attached Observations are the heart of the Pheontyping database. The information stored in the Experiment and supporting tables describes the design of the Field Trials and/or the results of statistical analysis. The Germplasm included in an Experiment are indicated by links to the Stock table.

   The Experiment table supports polymorphic behavior in the IBDB Phenotyping Database. It will store raw observations as well as summary statistics, means and even study templates. The different types are distinguished in the type column.

   Types are their usage
   a. Type = template, linked by proj_id, no links to the stock table. The purpose is to generate a "blank fieldbook" and capture properties of the experiment at the study level (e.g. irrigation)
   b. Type = location, linked by proj_id, no links to stock table. Usage of properties is similar to study above.
   c. Type = plot. A field trial. Linked by proj_id and links into the stock table. The properties describe the plot. Raw observations are linked from here in the phenotype table.

d. Type = means. Linked by proj_id, links to the stock table, properties determined by what is being tracked. Means are linked from here in the phenotype table.
e. Type = summary stats. Linked by proj_id, links to the stock table, properties similar to plot.

Properties of the Experiments are dependent on the type of experiment. Some properties can occur in multiple experiment types. For example, a location may be irrigated or have only certain plots which are irrigated. In this example, the irrigation property would be tracked at the location and plot level respectively.

Example properties for Plots include the physical layout of Field Trials (e.g. Plot, SubPlot, Rep), for locations conditions of the trial (e.g. fertilized, irrigated), and for templates the study level information (e.g. Principal Investigator, Objectives). The type_id for a property maps it back to the appropriate key in the Project (Study) properties table.

Refer to mappings in Table 1: Example Project Properties.

TABLE 2 EXPERIMENT PROPERTY ENTRY SHOWING PRINCIPAL INVESTIGATOR

experiment_id	**type_id (links to "link to ontology" column in project_prop)**	**value**	**rank**
1	2	Arllet	1

5. **Phenotype table (Observations)**
Observations are recorded in the phenotype table. All rows are linked to the appropriate experiment row and type of observation is determined by type of experiment.

In the case of experiments of type Plot the observations in the phenotype table will be the directly observed raw data about plants in field trial (e.g. yield, plant height).

In the case of experiments of type Means or Summary Stats, the observations in the phenotype table will be the results of statistical analysis on the raw data from Plot experiments. These results will be stored rather than computed in the database due to the complexity of the analysis even for simple means calculations.

The type of observation is linked back to the Project Property table using the attr_id column. In the current understanding of the business rules, observable_id and attr_id are identical. The team will perform further analysis to determine if both columns are required.

Below is a sample entry from the Phenotype table with Yield 10.3, Plant Height 80 and BLB Resistance score 3. Units for each of the measures (e.g. cm) are also stored in the project_properties table but are deleted for brevity.

The corresponding entries from the project property table are also included for clarity. Note in this example the type_id is shown rather than the text for the ontology type.

TABLE 3 PHENOTYPE DATA AND PROJECT PROPERTIES

Observable_id	Attr_id	Value
12	12	10.3
13	13	80
15	15	3

Project_id	Type_id	Value	Rank
2	1004	YIELD	26
2	1005	Grain yield	26
2	1006	12	26
2	1004	PHT	27
2	1005	Plant height	27
2	1006	13	27
2	1004	BLB	28
2	1005	BLB resistance score	28
2	1006	15	28

6. CV_term table (Ontology)

The cv_term table maps to the Ontology module of the Logical Data Model. The Ontology module contains the standard data dictionary for all terms in the Phenotyping Data Model. As described above, project_properties are used to map these terms to local equivalents.

Linking to the OMS system. cv_term replaces OMS_STDVARS and cv_term_prop replaces OMS_Classes in the existing IBDB.

VII. Appendix C – Breeding Problems

To further exercise the proposed model and finalize the list of business rules, the team will populate three breeding problems into the Chado schema. Full details of the problems and mappings will be provided as separate report according to the schedule provided in Next Steps.

Breeding Problem 1

1 Location, 12 subplots, multi-factorial trial

Breeding Problem 2

Multi-location, single year variety trial

Breeding Problem 3

Same as Breeding Problem 2 with some germplasm in common, e.g. cross-study, planning for next year trials

VIII. Appendix D – Schema Candidates

The team prioritized analysis of schema that were freely available by license and easily attainable. Commercial candidates listed in Table 2 could be evaluated at a future date if the open source candidates do not provide a good solution.

TABLE 4: BEST SCHEMA CANDIDATES USING CRITERIA

Name	Description	Comment
Chado with ND Extension	ND Extensions are Phenotyping system	Modular and extensible. ND applicable to domain, most of the rest not necessary. The Chado schema is freely distributed under the terms of the Artistic License (http://www.opensource.org/licenses/artistic-license.php) from GMOD (www.gmod.org). Have access to a Chado contributor, Lukas Mueller. Under analysis.
Germinate	Generic DB for	May be applicable. Available under GNU

	integration genotypic and phenotypic information for plant genetic resource collections	General Public License. Does not have the properties tables feature of Chado. Under analysis.
Katmandoo	Bioscience DMS	May be applicable. Free with a restricted license. Under Analysis.

TABLE 5 ADDITIONAL CANDIDATES

Name	Description	Notes
Grin Global	Inventory System	Commercial, does not apply to domain
Agrobase	Commercial Application	Commercial
Prism	Commercial Application	Commercial
Maize Finder	Maize Specific Application	Too specific to a single problem. NOTE: Functionality should be input to final choice of model
DMS/MDMS	Current system	Performance problems, non-modular approach to domain.

IX. Appendix E – Hardware and Software Baseline

The baseline configuration for running the tools and database for phenotypic data management.

- Laptop, no internet connection necessary
- 2G of Physical RAM, recommended 4G

- Pentium I5 or equivalent
- Windows 7 or higher
- HDD free space, 2G Program and 8G Data

X. Appendix F – Team

Name	Affiliation
Graham McLaren	GCP
Xavier Delannay	GCP
Brent Whitney	Efficio
Lynn Rogala	Efficio
François Schiettecatte	Efficio
Lukas Mueller	BTI
Hamer Paschal	GCP