**Phenotyping Data Management System for IBDB V3**

**INTRODUCTION**

The Data Management System (DMS) is the IBDB component of that manages germplasm characterization and evaluation data for genetic resources and crop improvement projects. DMS links these data to germplasm and pedigree information in the Genealogy Management System (GMS), to location information in the Location Management Module (LMM) and to information on genes, markers and alleles in the Genotyping Data Management System (GDMS) as well as providing links to other specialized data sources. DMS also allows integration of data from different characterization and evaluation studies, thus permitting a broad range of queries across trials or types of variables.

The Functions of the DMS are to:

1. Store and manage documented and structured data phenotyping data from germplasm characterization and evaluation studies as commonly conducted in crop improvement programs
2. Link data to specialised data sources such as GMS, soil and climate databases and
3. Facilitate queries, searches and data extraction across studies according to structured criteria for data selection as required by plnat breeders.

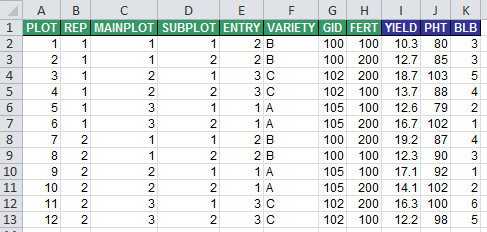
All types of phenotyping data will be accommodated in DMS including raw observed data, derived data, means data and summary statistics. Data may have numeric values or character values or categorical class values. For example, observations on disease resistance or nutrient efficiency of a genotype can be numerical measurements, scored or calculated indices or character data.

As a general principle, any data that are routinely represented in field books or laboratory books or spread sheets will be accommodated in DMS. It will handle data and documentation on the basis of individual phenotyping studies which may contain data from many environments (trial instances), and from different sampling levels – study level, environment level, plot level, sample level. Integration across studies will be facilitated by the use of controlled vocabularies and ontologies managed by the Ontology Management System (OMS)

**Structure of Phenotyping Data**

In order to clarify the definition of entities in the DMS data model we will consider data from a fictional split-plot field experiment. Although this type of experiment is not commonly used in crop improvement it does contain all the elements of the variety of experimental designs which are used and so serves as a good motivational model. Data sets are typically arranged in columns as in Table 1, we call the columns VARIABLES since each row may record a different value for each variable. The first seven variables define the source and context of the data. We refer to such variables as LABELS. The last three variables contain measured data. We call these data variables VARIATES, there are usually several variates in a data set. In general we can think of the LABELS as variables for which we know the values before we do the experiement, and VARIATES, the ones we measure during the experiment.

Table 1. Serial Spread-sheet Representation for a Split-plot experiment from Study S9801



STUDIES

A STUDY is the basic, reportable unit of research, it is synonymous with the notions of experiment, nursery or trial. Since DMS must deal with any of these we will use the term study. A study may be characterized by a set of scientific objectives and testable hypotheses and results in the collection of one or more datasets similar to that in Table 1 or it may be simply a convenient package of research activities such as all the replicated field evaluation for a breeding program in a particular year. A study always has some metadata associated with it, such as its name, the PI, the institute, IP status and so on. These are variables, or more precisely labels which take a single value which applies to the whole study. We call them STUDY LABELS, and they often appear as headers to tables such as Table 1 above.

The division of data into sets is usually motivated by convenience, for example data collected from different sampling scales is most conveniently treated in different datasets. Similarly, data collected at different times or from different locations are also often treated as different data sets, although it is feasible and usually preferable to treat these divisions in a single dataset.

Each row in a dataset corresponds to an OBSERVATION UNIT of the study. Values of STUDY labels apply to all the observation units in a study (from any dataset in the study).

ANNOTATION OF VARIABLES

Variables are named and described freely by users, but consistently within each study. However they are annotated by terms from three controlled vocabularies:

* The PROPERTY which describes the context of the sampling unit and experiemental material, if the variable is a label, or the trait being measured if it is a variate,
* the METHOD which describes how the PROPERTY is applied or the protocol by which a variate is measured and
* the SCALE which describes the units in which the label levels or variate values are recorded.

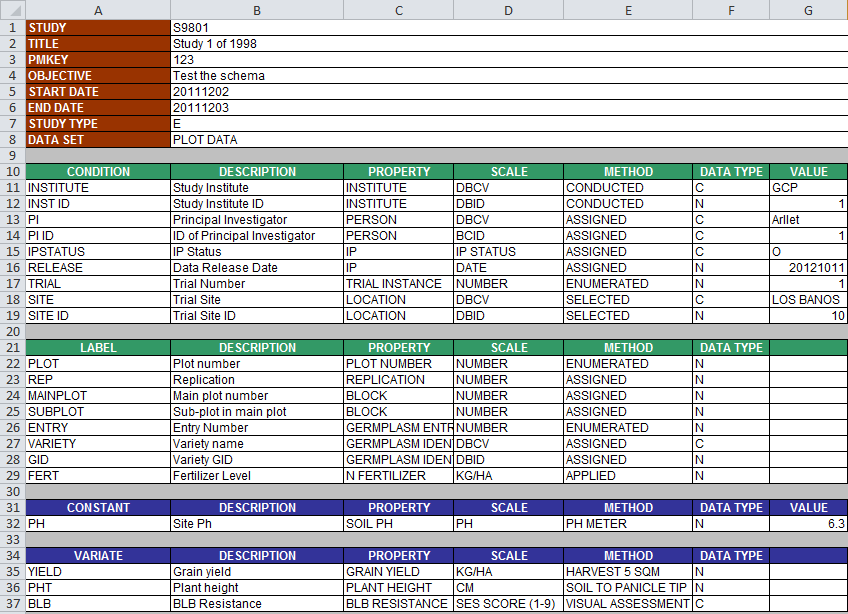
These three controlled vocabularies are terms in the Crop Ontology, and together they define the variables in the database. Every variable in the database is annotated by a combination of PROPERTY, METHOD and SCALE terms, and every unique combination of these terms occurring in the database defines a STANDARD VARIABLE. STANDARD VARIABLES are given standard names and descriptions in the ontology, but are referred to locally (within a study) by local names and descriptions assigned by the researcher. STANDARD VARIABLES link data across studies, and sets of STANDARD VARIABLES, for example those with the same property, or those with the same PROPERTY and METHOD or PROPERTY and SCALE link data about the same property across studies. All values of a particular STANDARD VARIABLE should be of the same data type. At present three types are being considered numeric, character and database IDs (links to records in other database modules such as GIDs). Variables can also be categorical in which case they can only take on values from a defined set of VALID VALUES. This range can be extended to cover other types such as binary, picture, URL, file link or other object.

LABELS and LEVELS

LABELS are classifying variables in a study which take values from finite sets of discrete LEVELS. These levels document the source and context of the data by expressing the conditions under which the data were collected or derived. For example, the names of treatments or design structures applying to the unit or units from which the data are recorded, or conditions such as the time and location of measurement. These LABELS are usually listed in columns in the data set as in Table 1. The Study Name will be treated in the data model as a LABEL with exactly one level. Hence, every study has at least one LABEL.

In phenotyping experiments we can identify four groups of labels which describe different parts of the study – STUDY labels, LOCATION (or environment) labels, ENTRY (or germplasm) labels, and FIELD TRIAL (or design) labels. In the example in Table 2, rows 1 to 8 and 11 to 16 describe STUDY labels, rows 17-19 describe LOCATION labels, rows 26-28 describe ENTRY labels, and rows 22 to 25 and 29 describe FIELD TRIAL labels. Labels listed in the CONDITION section have only one level or value for the particular data table annotated by the description sheet. Labels listed in the LABEL section have multiple levels and correspond to columns in the observation table (Table 1). Combinations of one level from each label define the observation units – rows of a spreadsheet as shown in Table 1.

Table 2. Description Sheet showing the annotation of variables for the data shown in Table 1



The role of a variable being a CONDITION or and LABEL is dependent on the scope of the data table. For example if data were collected from several sites, then the complete data set including data from all sites would have to have a label column indicating the location from where the data for that row was collected so SITE would be a LABEL. IF you only show a part of the dataset coming from one location then SITE is a CONDITION for that data table.

VARIATES AND VALUES

VARIATES are the variables which contain the data observed in the experiment – the phenotypic data. They usually appear as columns in the data table as YIELD, PHT and BLB in Table 1. Variates which have only one value pertaining to all observation units in the data table are called CONSTANTS.

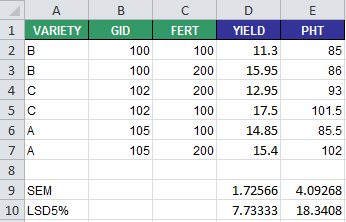
As discussed for CONDITIONs above, the status of CONSTANT depends on the data shown. If data in Table 1 were for two locations then PH would have to be represented as a VARIATE in the data table (ie a column) and similarly, SITE would be a LABEL in the data table.

OBSERVATION UNITS

Data sources such as field objects or sampling units are identified by combinations of levels of design or sampling LABELS. In Table 1, the LABELS REP, MAINPLOT and SUBPLOT are all design LABELS and combinations of one level from each identify physical sub-plots in the study. Other LABELS define the context of the data, in experiments these are called treatment LABELS. Combinations of one level from each treatment LABEL define the treatments which are applied to field objects. In Table 1 VARIETY and FERT are treatment LABELS.

Data values such as treatment means, as in Table 3, are associated with level combinations of treatment LABELS which do not correspond to field objects but which can be thought of as data sources. Both types of data sources, field objects and treatment combinations, are referred to as OBSERVATION UNITS.

Table 3. Least Squares treatment means for data in Table 1 – Study S9801.



Hence OBSERVATION UNITS are conceptually equivalent to rows in a serially structured spreadsheet, they are the real or conceptual data sources in a study and they are annotated by distinct level combinations of one or more LABELS. Not all LABELS in a study need to be involved in this indexing for every OBSERVATION UNIT. However, STUDY LABELS, with their single levels, are involved in indexing every OBSERVATION UNIT in the Study. Hence OBSERVATION UNITS belong to unique studies. Every study has a STUDY UNIT which is the single observation unit indexed by the level of the STUDY LABEL alone.

DATASETS

All observation units in a study which have the same labels form a DATASET. We can say that a DATASET is defined by the different level-combinations of a subset of LABELS from the STUDY. The OBSERVATIONS UNITS of the PLOT DATA in Table 1 are indexed by PLOT, REP, MAINPLOT, SUBPLOT, ENTRY, VARIETY, GID and FERT and for the TREATMENT MEANS in Table 2, ENTRY and FERT will define all the OBSERVATION UNITS, but we would like to carry over the other labels - ENTRY, VARIETY and GID form the plot data table, as well as all the STUDY LABELS from the study record to the means table in order to complete the annotation of the means.

The summary statistics SEM and LSD5% are also variables and belong to a TRIAL ENVIRONMENT dataset indexed by the STUDY LABELS and the SITE or ENVIRONMENT labels

**LOGICAL DATA MODEL FOR PHENOTYPING DATA FROM FIELD TRIALS**

Using the above generic definitions of the components of data from a phenotypic study we can consider a more precise model for handling data from a field trial. There are five key elements of phenotyping data from field trials which need to be captured in a logical data model. These are:

* The STUDY AND DATASET INFORMATION component which records global contextual information about the experiment such as who conducted it, when, why, and who owns the data. It also models the high level structure of the experiment by describing the datasets that are part of the study for example, data collected about the trial environment(s), data collected on sub-samples, plot data, means and summary data. Each of these datasets is described in terms of the variables that occur in them, which ones are labels giving design and context, and which are variates containing observations made during the experiment. The actual values of these dataset variables are managed by other components of the model depending on their type.
* The TRIAL ENVIRONMENT component manages all data values describing the environments observed in the study including georeference information, place names, growing environments, and overall management practices (non-treatment factors). This component also links to the Location Management Module (LMM) of the IBDB.
* The GERMPLASM ENTRY component manages all label values describing the germplasm entries in the experiment including local and global identifiers, names, sources and roles (check or test lines) of the entries. It links to the Germplasm Management System of (GMS) IBDB where unique identification, global nomenclature, ownership and pedigree information is stored.
* The TRIAL DESIGN component manages the treatment and sampling design and structure of the datasets in the study. It enumerates all the observation units in the study and describes their treatment and sampling context in terms of levels of labels describing such features as replication, block, fertilizer treatment etc. the observation units inherit global information about the study, information about the location and information about the germplasm entries by linkages to the relevant components of the model.
* The OBSERVATION component manages the values of the variates for each dataset.
* The ONTOLOGY component has two parts, a set of structural terms which define the above components of a trial within the schema and a variable part which defines the variables in a trail and link them across different studies within a crop.

Figure 1a. Key elements of the Logical Data Model for Penotyping Data from Field Trials

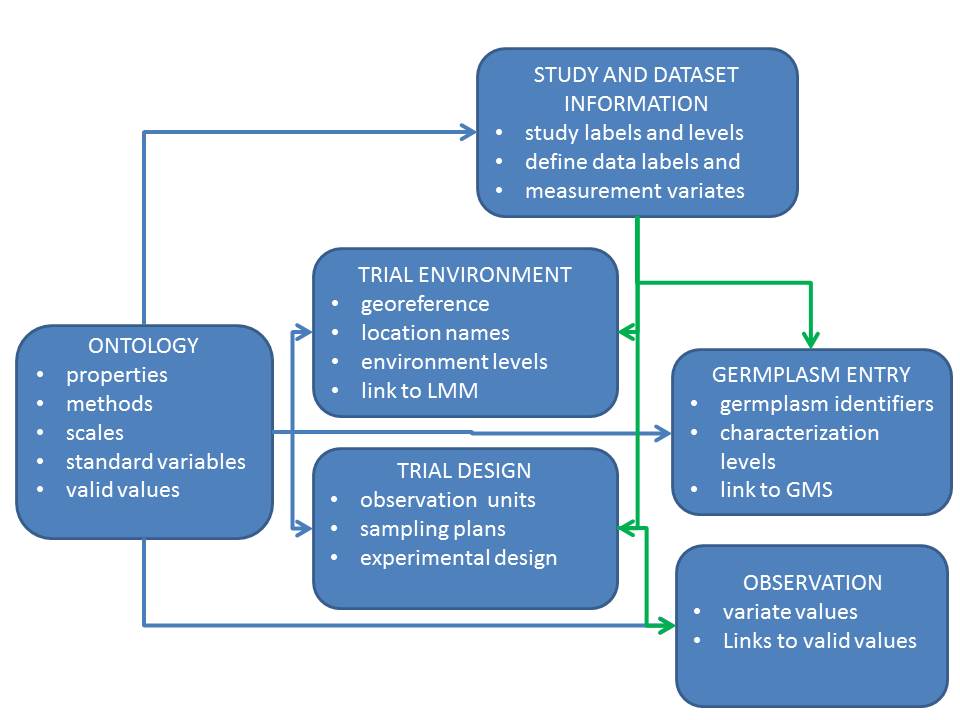
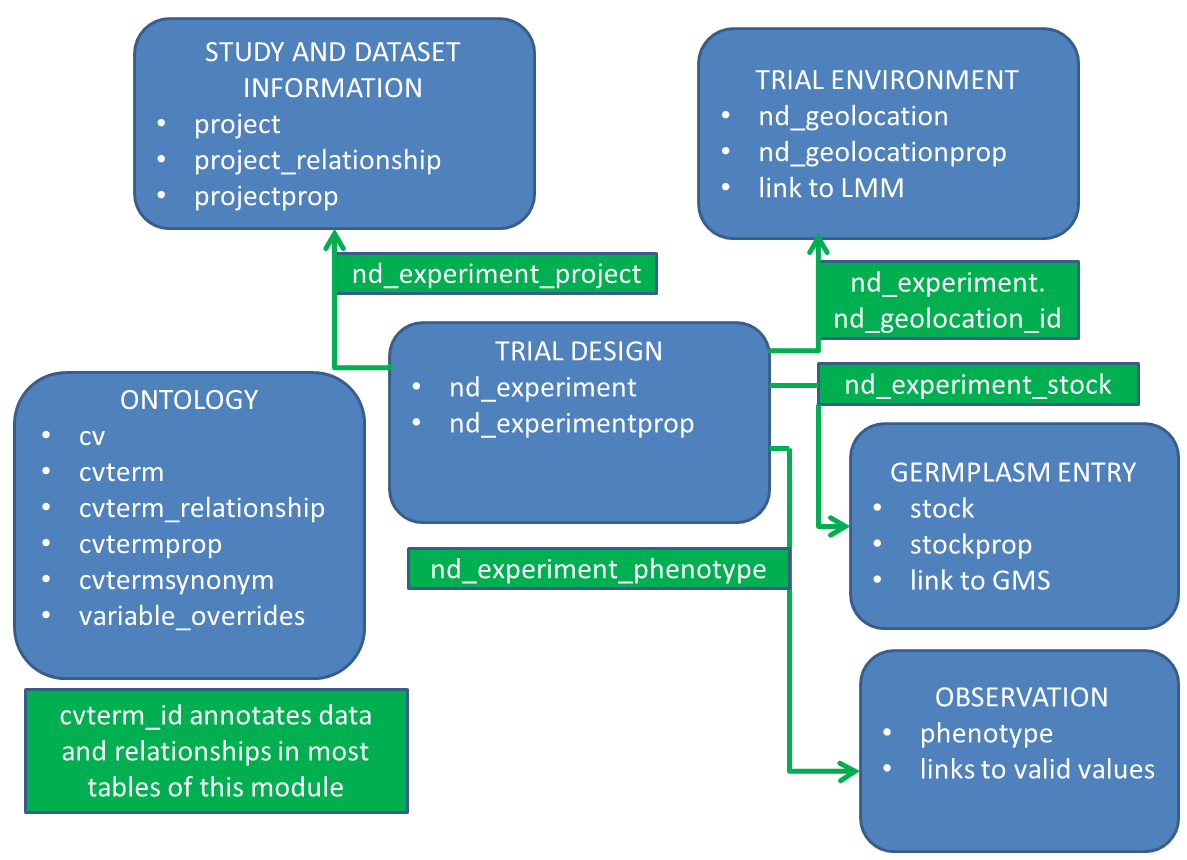


Figure 1b. Mapping of elements of the Logical Data Model to the CHADO ND schema



**BUSINESS RULES FOR USING THE CHADO NATURAL DIVERSITY MODULE TO MANAGE PHENOTYPING DATA FROM FIELD TRIALS**

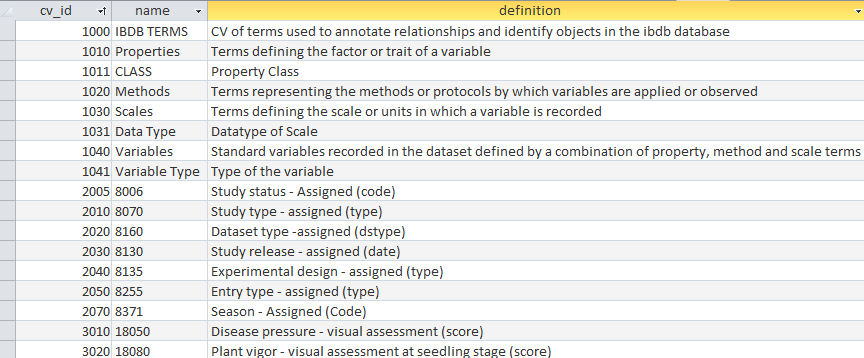
**The Ontology Management System**

Two ontologies are used to annotate different elements of phenotyping data so that they can be consistently stored in the CHADO ND schema and compatible data can be integrated across different phenotyping studies.

* Firstly, the IBDB Structure ontology has one controlled vocabulary, the IBDB TERMS CV which has terms defining properties of different components of the data model and relationships between them. This essentially extends the CHADO ND schema.
* Secondly, the Crop Ontology defines the variables which describe the context of the phenotyping experiment and which record the observations from the experiment. The Crop Ontology comprises several controlled vocabularies (CV).

Each of controlled vocabulary is defined in the ***cv*** table and the terms for each cv are in the ***cvterm*** table.

Figure 2. Controlled vocabularies for Phenotyping Data from Field Trials



**The IBDB Structure Ontology**

The IBDB Structure Ontology comprises a single controlled vocabulary, the IBDB TERMS CV (cv\_id 1000). Terms in this cv belong to one of five classes:

* IBDB structure,
* IBDB element,
* IBDB class,
* IBDB data type and
* UI Label.

The relation of belonging to a class is recorded by a record in the cvterm\_relationship table with type\_id=1225 (is a). Terms in the class IBDB structure describe entities and relationships used in the schema. Terms in IBDB class group the IBDB terms in a hierarchy for easy browsing. IBDB TERMS which are class terms also have a ‘type’ relationship (1105) to term 1090 ‘Class’ and an ‘is a’ relationship (1225) to a parental class. Terms in IBDB element indicate the component of the logical schema to which a STANDARD VARIABLE belongs, and where its values are stored. IBDB Terms with class UI Label are mapped to features in the user interface and allow customization of labels for these features by changing the names of the terms.

**The Crop Ontology**

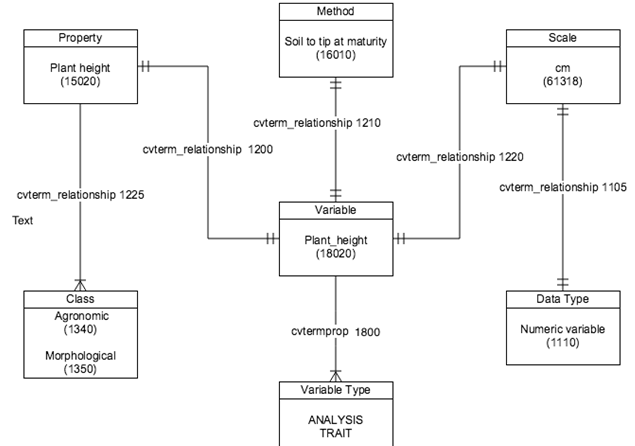
The Crop Ontology comprises the following controlled vocabularies and relationships amongst them.

* The VARIABLES CV has terms defining what is being recorded, how it is recorded and what scale or units it is recorded in for all variables (labels and variates) involved in a phenotyping study. They are the terms used to annotate each data point in a study and they are defined by unique combinations of one property term, one method term and one scale term from the other controlled vocabularies in the Crop Ontology.
* The PROPERTIES CV has terms defining nature of the variables in a phenotyping experiment.
* The CLASS CV has terms defining a classification of PROPERTIES for grouping and navigation. Each PROPERTY term belongs to one or more classes.
* The METHODS CV has terms describing the protocols by which those properties are applied or measured in phenotyping experiments.
* The SCALES CV has terms describing the scales or units in which the values of the properties are recorded.
* The DATA TYPE CV has terms defining the type of data associated with each SCALE.
* Each scale which has the categorical data type has a VALID VALUE CV with terms describing the valid values for the scale and their meanings.
* The VARIABLE TYPE CV specifies a type of variable indicating its role in the field trial workflow. Each variable has one or more VARIABLE TYPEs.

The relationships between these CV which ultimately define the variables are recorded as relationships in the cvterm\_relationship table or as properties in the cvtermprop table as indicated in the ontology data model in Figure 3.

The PROPERTIES CV has terms which define the nature of each variable in a phenotyping study. Essentially they define the ‘thing’ being recorded, applied or measured such as the altitude of a test site, the amount of nitrogen fertilizer applied or the grain yield from a plot of corn. PROPERTIES terms do not indicate how the values of the variables arise or in what units they are expressed. Each PROPERTY term has an ‘is a’ (1225) relationship to one or more terms in the CLASS CV. The CLASS CV provides a tree classification of PROPERTY terms for ease of navigation and searching. The top division in the tree is into PROPERTIES which describe the context and structure of the phenotyping study – the Crop Research Ontology on the one hand and PROPERTIES which are measured in the studies – The Crop Trait Ontology.

Figure 3. The Ontology Data Model



The METHODS CV has terms which describe how the properties of a particular variable are applied, observed or measured to arrive at the values recorded in the study. The SCALES CV has terms describing the scales or units in which the values of the properties are recorded. Each term is related to a term in the DATA TYPE CV by the ‘has data type’ (type\_id=1105) relationship.

Table 4. The DATA TYPE CV

| **cvterm\_id** | **Name** | **Definition** |
| --- | --- | --- |
| 1110 | Numeric variable | Variable with numeric values either continuous or integer |
| 1117 | Date variable | Date - numeric value in format yyyymmdd with least significant parts set to zero acording to precision |
| 1118 | Numeric DBID variable | Integer database ID (may be negative) |
| 1120 | Character variable | Variable with character values |
| 1130 | Categorical variable | Variable with discrete class values |
| 1131 | Person | Datatype which represents person |
| 1132 | Location | Datatype which represents location |
| 1133 | Study | Datatype which represents study |
| 1134 | Dataset | Datatype which represents dataset |
| 1135 | Germplasm List | Datatype which represents germplasm list |
| 1136 | Breeding Method | Datatype which represents breeding method |

Each scale which has the categorical data type has a VALID VALUE CV with terms describing the valid values for the scale and their meanings.

The CLASS CV has terms defining a classification of PROPERTIES for grouping and navigation. Each PROPERTY term belongs to one or more classes.

Table 5. Terms of the CLASS CV arranged into their tree structure.

**Cvterm\_id Term name**

* 1045 **Crop research ontology**
* 1050 Study condition
* 1055 Dataset Condition
* 1080 Trial environment
* 1280 Abiotic condition
* 1300 Site condition
* 1310 Soil condition
* 1320 Climatic condition
* 1290 Biotic condition
* 1085 Trial management
* 1086 Variate condition
* 1087 Germplasm
* 1100 Trial Design
* 1260 Seed storage
* 1270 Breeding process
* 1321 Molecular property
* 1460 Breeding methods
* 1470 Generative breeding methods
* 1560 Crossing methods
* 1570 Mutation methods
* 1580 Cytogenetic manipulation
* 1590 Genetic Modification
* 1480 Derivative breeding methods
* 1490 Bulking breeding methods
* 1510 Non-bulking breeding methods
* 1520 Maintenance methods
* 1530 Seed increase methods
* 1540 Seed acquisation methods
* 1550 Cultivar formation methods
* 1330 **Crop trait ontology**
* 1340 Agronomic
* 1430 Yield component
* 1440 Phenology
* 1450 Post harvest
* 1345 Physiological
* 1350 Morphological
* 1360 Biotic stress
* 1390 Disease resistance
* 1400 Insect and pest resistance
* 1364 Biochemical trait
* 1370 Quality
* 1380 Passport
* 1410 Abiotic stress
* 8580 General

Each term in the CLASS CV has a ‘type’ relationship (1105) to term 1090 ‘Class’ and an ‘is a’ relationship (1225) to a parental class. This is how the tree structure is recorded.

Table 6. VARIABLE TYPES CV

| **cvterm\_id** | **Name** | **definition** | **Stored in** |
| --- | --- | --- | --- |
| 1805 | Study Detail  **Examples:**  study name, study objective, PI name, Institution, project code | Administrative details to be tracked per study  **Displayed at:**  Nursery Manager > Nursery Settings > Management Details  and  Trial Manager > Settings > Trial Settings | ***projectprop*** table.  Except for Study Name (8005) and Study Title (8007) and Dataset Name (8150) and Dataset Title (8155) which are stored in ***project.name*** and ***project.description***. Otherwise values are stored as text in ***projectprop.value*** unless the variable has categorical scale in which case the ***cvterm\_id*** of the valid value is stored as a string in ***projectprop.value***. |
| 1806 | Environment Detail  **Examples:**  site name, site coordinates, collaborator name, institution, experimental design  Examples:  plot size, harvest date, planting date, irrigation date, irrigation amount, fertilizer use, fertilization date, pesticide use | Administrative details to be tracked per environment  **Displayed at:**  Trial Manager > Environments > Environment Details  and  Nursery Manager > Nursery Settings > Management Details | ***geolocationprop*** table linking to ***nd\_geolocation*** records for each trial environment  Except for Trial Instance (8170) which is stored in ***nd\_geolocation.description*** and Site lat, long, datum and altitude (8191,8192,8193,8194) values which are also stored in the ***nd\_geolocation*** table. Otherwise values are stored as text in ***nd\_geolocationprop.value*** unless the variable has categorical scale in which case the ***cvterm\_id*** of the valid value is stored as a string in ***nd\_geolocationprop.value***. |
| 1802 | Trial Condition  Examples:  soil\_pH, rainfall, weather, environmenal comments | Observations made of conditions in an individual environment involved in a trial  **Displayed at:**  Trial Manager > Environments > Trial conditions  and  Nursery Manager > Nursery Settings > Nursery conditions | ***phenotype*** table linking to nd\_experiment record of the Environment dataset  Variable values are stored in phenotype.value unless the data type is categorical in which case the cvterm\_id of the valid value is stored in phenotype.cvalue\_id. |
| 1803 | Nursery Condition | Observations made of conditions in a nursery  **Displayed at:**  Trial Manager > Environments > Trial conditions  and  Nursery Manager > Nursery Settings > Nursery conditions | ***phenotype*** table  Variable values are stored in phenotype.value unless the data type is categorical in which case the cvterm\_id of the valid value is stored in phenotype.cvalue\_id. |
| 1804 | Germplasm Descriptor | Information to be recorded about each germplasm in a study  **Displayed at:**  Nursery Manager > Nursery Settings > Germplasm Descriptor  and  Trial Manager > Germplasm > Germplasm Descriptor | ***stockprop*** table  Except for variables ENTRY\_NO (8230), GID (8240), DESIGNATION (8250) and ENTRY\_CODE (8300) which are stored in ***stock.uniquename, stock.dbxref\_id, stock.name and stock.value*** respectively and variable Stock\_ID (8269) which is stored in ***ims\_transaction.inventory\_id.*** |
| 1807 | Selection Method  Examples:  NPSEL, | How material is chosen for advancing to the next generation  **Displayed at:**  Nursery Manager > Nursery Settings > Selection | ***phenotype*** table linking to the ***nd\_experiement*** record for the whole environment or for each plot if plots differ.  Except for variables with property Inventory Amount for which values are stored in ***ims\_transaction.trnqty.*** |
| 1810 | Experimental Design  Examples:  plot, subplot, block, sub-block, row, column, repetition | Design to be applied to experiments  **Displayed at:**  These variables are selected by default in the measurements tab of the Nursery Manager, and according to the chosen design in the Design tab of the Trail Manager | ndexperimentprop |
|  | Sampling Design  Examples:  Sample no, quadrat, plant no, obs\_time | Sampling design for multiple measurements from a plot. | ndexperimentprop |
| 1809 | Treatment Factor  Examples:  N\_Fert  N\_fert\_applied | Treatments to be applied to members of a trial  **Displayed at:**  Trial Manager > Treatment Factors | ndexperimentprop |
| 1808 | Trait  Examples:  plant height, seed color, rust incidence, seed protein content | Characteristics of a germplasm to be recorded during a study  **Displayed at:**  Nursery Manager > Nursery Settings > Traits  and  Trial Manager > Measurements > Traits | ***phenotype*** table  Variable values are stored in phenotype.value unless the data type is categorical in which case the cvterm\_id of the valid value is stored in phenotype.cvalue\_id. |
|  | Derived Trait  Examples:  Grain\_Yield\_kgha,  ASI | Characteristics of a germplasm computed from one or more measured traits and trial conditions | ***phenotype*** table  Variable values are stored in phenotype.value unless the data type is categorical in which case the cvterm\_id of the valid value is stored in phenotype.cvalue\_id. |
| 1801 | Analysis  **Examples:**  max, min, median, LSD, heritability, error | Variable to be used only in analysis (for example derived variables) | ***phenotype*** table  Variable values are stored in phenotype.value unless the data type is categorical in which case the cvterm\_id of the valid value is stored in phenotype.cvalue\_id. |

Terms in the VARIABLES CV are related to one term in the PROPERTY CV, one METHOD term, and one SCALE term. Each VARIABLE term also has a VARIABLE TYPE which is recorded as a property of the variable in the cvtermprop table with type\_id-1800 and value equal to the name of the variable type (not the id).

For example term 8250 DESIGNATION has relationships shown in Table 7 and the property shown in Table 8

Table 7: Relationships and property for VARIABLE 8250 (DESIGNATION)

| **cvterm\_relationship** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **cvterm\_relationship\_id** | | **type\_id** | | **subject\_id** | | | **object\_id** |
| 8240 | | 1200 | | 8250 | | | 2205 |
| 1530 | | 1225 | | 2205 | | | 1087 |
| 8242 | | 1210 | | 8250 | | | 4030 |
| 8244 | | 1220 | | 8250 | | | 1908 |
| 22053 | | 1105 | | 1908 | | | 1135 |
| **cvtermprop** | | | | | | | | |
| **cvtermprop\_id** | | **cvterm\_id** | | **type\_id** | **value** | | | **rank** |
| 20419 | | 8250 | | 1800 | Germplasm Descriptor | | | 0 |

Which should be read as follows:

Subject 8250 (DESIGNATION) has PROPERTY (type\_id 1200) 2205 (GERMPLASM ID), (which ‘is a’ (type\_id 1225) GERMPLASM TERM. It also has METHOD (type\_id 1210) 4030 (APPLIED) and has SCALE (type\_id 1220) 1908 (GERMPLASM NAME) which has DATA TYPE (type\_id 1105) 1135 (GERMPLASM LIST). It has VARIABLE TYPE Germplasm Descriptor.

The VALID VALUE CVS

Each SCALE with Data Type 1130 (CATEGORICAL VARIABLE) spawns a CV of VALID VALUES. These cvs contain the valid values and their interpretation for the categorical variable to which this scale applies. They are named in the ***cv*** table by the string value of the ***cvterm\_id*** from the VARIABLES CV for the categorical variable to which they belong (although this is not the link which is described below), and they are described in the ***cv*** table with the description of the categorical variable. For example the valid values for the scale 25710 (Type of EXPT\_DESIGN) are shown in Figure 4. This scale is referenced by VARIABLE 8135 EXPERIMENTAL DESIGN - ASSIGNED (TYPE). Although it is possible to use a categorical scale for more than one VARIABLE this is not recommended since values could be added or changed with reference to one variable thereby corrupting data annotation for another variable.

Figure 4. The cv definition and valid value terms for scale 25710 (Type of EXPT\_DESIGN)

| **cv** | | | |
| --- | --- | --- | --- |
| **cv\_id** | **name** | **definition** | |
| 2040 | 8135 | Experimental design - assigned (type) | |
| **cvterm** | | | |
| **cvterm\_id** | **cv\_id** | **name** | **definition** |
| 10055 | 2040 | UNREP | Unreplicated demonstration |
| 10100 | 2040 | CRD | Completely randomized design |
| 10110 | 2040 | RCBD | Randomized complete block design |
| 10120 | 2040 | Alpha | Alpha lattice |
| 10130 | 2040 | RIBD | Resolvable incomplete block design |
| 10140 | 2040 | NRIBD | Non resolvable incomplete block design |
| 10145 | 2040 | RRCD | Resolvable row-column design |
| 10150 | 2040 | NRRCD | Non resolvable row-column design |
| 10160 | 2040 | Augmented | Augmented design |
| 10165 | 2040 | SP | Split-plot design |
| 10166 | 2040 | RIBDL | Resolvable Incomplete Block Design (Latinized) |
| 10167 | 2040 | RRCDL | Resolvable Row-and-Column Design (Latinized) |

Each VALID Value is related to its scale by a relationship with type ‘has value’ (type\_id=1109). For example SCALE 2571 has the following cvterm\_relationship record for value UNREP (cvterm\_id=10055):

| **cvterm\_relationship** | | | |
| --- | --- | --- | --- |
| **cvterm\_relationship\_id** | **type\_id** | **subject\_id** | **object\_id** |
| 10055 | 1190 | 25710 | 10055 |

Categorical variables sometimes require an ordering for their values (either because there is an intrinsic ordering eg Low, Medium, High, or because it makes sense to present them to users (in pick lists for example) in a certain order. The default order is the alphabetical order of ***cvterm.name*** (with numbers treated in character order). If a different ordering is required each term should have a property in the ***cvtermprop*** table of ***type\_id*** order (IBDB TERMS ***cvterm\_id***=1420) with the numerical sequence order for that term as its ***value***.

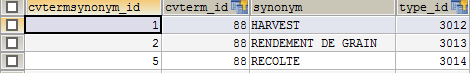
SCALES of type 1110 (NUMERIC VARIABLES) may have MINIMUM and MAXIMUM allowable values specified in the ***cvtermprop*** table as shown in Table 9 for variable SOIL PH.

Table 9: MINIMUM and MAXIMUM allowable values for a numeric variable in the CRO

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| cvtermprop\_id | cvterm\_id | type\_id | value | rank |
| 8000 | 8270 | 1113 | 1 | 0 |
| 8010 | 8270 | 1115 | 14 | 0 |

Synonyms and foreign language names and descriptions for terms are stored in the ***cvtermsynonym*** table.

Figure 4. Synonyms and foreign language names for controlled vocabulary terms



The LINKS TO [www.CropOntology.org](http://www.CropOntology.org)

The Crop ontologies in IBDB are supposed to be linked (and synchronized) with the terms on [www.CropOntology.org](http://www.CropOntology.org). The Term IDs from that site are carried as properties of the corresponding cvterms in the cvtermprop table with type\_id=1226 (Crop ontology term ID).

**Program Customizations**

The variable\_overrides table is used to store aliases for variable names as well as valid ranges for continuous variates which are local to particular programs.

**The Data Management System**

STUDY AND DATASET INFORMATION

We are concerned with phenotyping STUDIES which are generally called trials or nurseries in the plant research context. We interpret trials as studies which include some form of replication, involve some form of randomized design and may be implemented in multiple instances (environments). Nurseries do not generally contain replication (although the treatment list can have repeated entries) do not involve randomized designs and are constrained to a single instance. By way of an example for the following description, the trial described in Tabel1 1 and 2 was loaded into the BMS using the ‘use my own design’ function of the Trial Manager (Annex 1). The resulting tables and figures are described and annotated below.

Data for each STUDY is managed as a series of related DATASETS. The nature, structure and relationships between these datasets are captured using the CHADO ***project*** tables. CHADO projects can be folders (and sub-folders), studies or datasets. The names and descriptions of folders, studies and datasets must be unique and are stored in the ***project*** table. Studies are organized into hierarchical structures of folders. Folders can contain sub folders or studies themselves. Studies contain DATASETS. The structure of folders, sub-folders, studies and datasets in managed through the ***project\_relationship*** table using the relationships “is a sub-folder of” . (***type\_id***=1140), ‘is a study in’ (***type\_id***=1145) and “is a dataset of”,(***type\_id***=1150). (The ***type\_ids*** take values which are ***cvterm\_ids*** for the relationships). Folders do not contain information (other than their names and descriptions in the ***project*** table) but they group sub-folders and studies for easy browsing. An example of the ***project*** and ***project\_relationship*** records for the study S9801 from Tables 1, 2 and 3 are shown in Figure 5.

Figure 5. The ***project*** and ***project\_relationship*** records for study S9801

| **project** | | | |
| --- | --- | --- | --- |
| **project\_id** | **name** | **description** | |
| 1 | STUDIES | Root study folder | |
| 25007 | S9801 | Study 1 of 1998 | |
| 25008 | S9801-ENVIRONMENT | S9801-ENVIRONMENT | |
| 25009 | S9801-PLOTDATA | S9801-PLOTDATA | |
| 25010 | S9801-MEANS | S9801-MEANS | |
| **project\_relationship** | | | |
| **project\_relationship\_id** | **subject\_project\_id** | **object\_project\_id** | **type\_id** |
| 25007 | 25007 | 1 | 1145 |
| 25008 | 25008 | 25007 | 1150 |
| 25009 | 25009 | 25007 | 1150 |
| 25010 | 25010 | 25007 | 1150 |

Every field trial and nursery has at least three project records, one for the study itself (S9801 in the example), one for environment data and one for plot data. The name of the study is freely chosen but must be unique, the names of the environment and plot datasets are created by appending ‘-ENVIRONMENT’ and ‘PLOTDATA’ respectively. A trial may also have means data saved after analysis and the name of the means dataset is derived by appending –MEANS to the study name as in the example in Figure 5.

Each item of information or data relating to a STUDY is identified/annotated with a variable ID from the VARIABLES ontology. For example the study name is identified with the variable 8005 - STUDY\_NAME which has property STUDY, method ASSIGNED and scale NAME. This identification is stored in the ***projectprop*** table where each variable in the associated study or dataset is represented by three records all having ***projectprop.project\_id*** equal to the ***project\_id*** of the study or dataset and the field ***projectprop.rank*** containing the rank order of the variable in the study or dataset. One of these records having ***projectprop.type\_id***=1070 specifies the ***cvterm\_id*** of the standard variable to which it belongs in the ***projectprop.value*** field. A second record with ***projectprop.type\_id***=1060 specifies a user-supplied description of the variable in the ***projectprop.value*** field and a third record has the VARIABLE TYPE of the variable (Table 6.) in the ***projectprop.type\_id*** field and a user-supplied name for the variable in the ***projectprop.value*** field. The user supplied names and descriptions may be different from the standard names and descriptions of the variable in the ***cvterm*** table. The ***cvterm\_id*** will link the variable across studies and to the standard names and descriptions as required for data integration. All ***projectprop*** records for the same variable have the same ***projectprop.rank*** in the ***projectprop*** table. The rank specifies the user-supplied order of the variables in the study.

The STUDY DATASET

The values of the study name (***cvterm\_id***=8005) and title (***cvterm\_id***=8007) variables are stored in the ***name*** and ***description*** fields of the study record in the ***project*** table.

The values of all ‘Study Detail’ variables (VARIABLE TYPE=1805) are stored in a fourth ***projectprop*** record with the same ***projectprop.rank*** value as the variable description records. If there are any study constant variates (observations or measurements having a single value for the whole study) they are similarly described in the projectprop table with project\_id of the whole study. The values however are stored in the ***phenotype*** table linked to a study observation unit (***nd\_experiment*** record linked to the study).

All other variables (labels and variates) in the study belong to one or more datasets and these are described in ***projectprop*** records, in the same way as study variables, for each dataset in which they occur. As for studies the levels of the dataset name and title labels, if present, are stored in the ***project*** table. Other labels which are dataset conditions (with VARIABLE TYPES 1805 or 1806) are stored in the ***projectprop*** table as with study labels. Values of all other variables in a dataset are stored in tables appropriate to their variable types, linked to observation units related to the dataset.

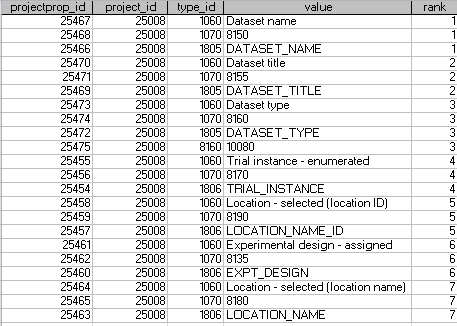
Figure 6. ***projectprop*** table records for the STYDY S9801 (project\_id=25007)



The ENVIRONMENT DATASET

In particular for trials and nurseries there must be an ENVIRONMENT dataset. The structural factor of this dataset is TRIAL\_INSTANCE with values 1,2, 3 … This variable has the single value 1 for single site trials and nurseries. The ***projectprop*** records for the labels in the S9801-ENVIRONMENT dataset are shown in Figure 7.

Figure 7. ***projectprop*** records for the Environment dataset of study S9801



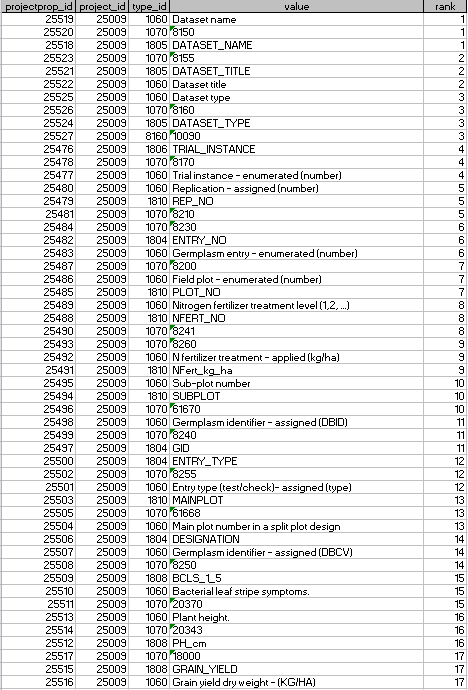
The environment dataset may contain variates having a single value for each environment with Variable Types 1802 (Trial Condition), 1803 (Nursery Condition) or 1801 (Analysis) they are similarly described in the projectprop table. The values however are stored in the ***phenotype*** table linked to an environment observation unit (***nd\_experiment*** record linked to the environment dataset). Often the Analysis variates are summary statistics from singe site analysis of a trial. For example projectprop records for the LSD of GRAIN\_YIELD would look like:



The PLOTDATA dataset.

Trials and nurseries must have a PLOTDATA dataset. The structural factors defining the observation units in the plot dataset are TRIAL\_INSTANCE, and PLOT\_NO and the main treatment factor is ENTRY\_NO.

Figure 8. ***projectprop*** trecords for the plot dataset of study S9801



The values of labels in datasets which are not Study Details and Environmental Details are stored in the TRIAL ENVIRONMENT, GERMPLASM ENTRY or TRIAL DESIGN components of the schema (Figure 1) according to their Variable Type (Table 6). Several cardinal variables (as noted in footnotes to Table 6) are stored in special fields of the CHADO tables associated with those components and other variables are stored in the property table associated with those components.

The values of the variates of studies or datasets are stored in the OBSERVATION component of the schema (in the ***phenotype*** table).

TRIAL ENVIRONMENT

The location/environment component of the logical data model manages information about the trial environment where an experiment is conducted. It uses the ***nd\_geolocation*** table and ***nd\_geolocationprop*** table to store all values of location labels. Information in these tables corresponds to actual physical locations where field trials or nurseries are conducted.

The ***nd\_geolocation.description*** field is used to store the value of factor with property TRIAL INSTANCE, method ENUMERATED and scale NUMBER. If the study comes from the BMS Trial Manager this is always present and contains a sequential number 1,2, …. But studies from other applications may have TRIAL INSTANCE factors with a different scale. If there is no TRIAL INSTANCE factor then simply assign sequential numbers to ***nd\_geolocation.description*** and put all LOCATION labels in the ***nd\_gelocationprop*** table. Georeference properties of the trial site are also stored in the ***nd\_geolocation*** table if available. All other labels of the trial location called Environment Details in Table 6. (eg site name, side code, site ID and Location ID, cooperator name etc.) are stored in the ***nd\_gelocationprop*** table linked to the appropriate ***nd\_geolocation*** record.

Global information about locations is stored in the Location Management Module of the database, outside the scope of the Phenotyping database, but linked via the Location ID label with property LOCATION, method ASSIGNED and scale location ID.

Other labels of the trial site which describe the management or environment of the trial (ie not treatments within the trial) have Variable Type ‘Trial Condition’ in Table 6. are also stored as properties of the trial environment in the nd\_***geolocationprop table***. (eg, irrigation, pesticides, season etc.)

Figure 8. nd\_geolocation table and properties SITE ID and SITE for the trial environment stored in the ***nd\_geolocationprop*** table.

| **nd\_geolocation** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **nd\_geolocation\_id** | **description** | | **latitude** | **longitude** | **geodetic\_datum** | **altitude** |
| 1 | 1 | |  |  |  |  |
| **nd\_geolocationprop** | | | | | | |
| **nd\_geolocationprop\_id** | | **nd\_geolocation\_id** | | **type\_id** | **value** | **rank** |
| 1 | | 1 | | 8180 | Los Banos - (LB) | 4 |
| 2 | | 1 | | 8135 | 10165\* | 3 |
| 3 | | 1 | | 8190 | 9016 | 2 |

\* 8135 is the categorical variable EXPT\_DESIGN and 10165 is the cvterm\_id for valid value SP (Split Plot).

GERMPLASM ENTRIES

The ***stock*** table maps to the Germplasm component 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.

The ***stock.uniquename*** field is used to store the label with property GERMPLASM ENTRY, method ENUMERATED and scale NUMBER (ENTRY\_NO, ***cvterm\_id***=8230) this is always present if the study comes from the Trial Manager, but if it is not present (study from another application) simply store a sequence number for the germplasm entries in the study in this field.

If there is a label with property GERMPLASM ID, method ASSIGNED and scale Germplasm id (GID, ***cvterm\_id***=8240) this contains the GID from GMS and they should be stored in the ***stock.dbxref\_id*** field. Else *null*. The ***stock*** table thus serves as a link to the information in the GMS via the ***stock.dbxref\_id*** field.

If there is a label with property GERMPLASM ID, method ASSIGNED and scale Germplasm name (DESIGNATION, ***cvterm\_id***=8250) it contains a germplasm name from the database and its levels should be stored in the ***stock.name*** field. Else *null*.

If there is another label with property GERMPLASM ID, method ASSIGNED and scale CODE (ENTRY\_CODE, ***cvterm\_id***=8300) it contains a study level entry code for the germplasm and its levels should be stored in the ***stock.value*** field and its ***cvterm\_id*** in the ***stock.type\_id*** field. Else both are *null*. Put 8300 (ENTRY\_CODE) in Type\_id.

Figure 9: Information on the entries in the experiment in the ***stock*** table.

| **stock** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **stock\_id** | **dbxref\_id** | **organism\_id** | **name** | **uniquename** | **value** | **description** | **type\_id** | **is\_obsolete** |
| 1 | 100 |  | B | 2 |  |  | 8300 | 0 |
| 2 | 102 |  | C | 3 |  |  | 8300 | 0 |
| 3 | 105 |  | A | 1 |  |  | 8300 | 0 |

The ***stockprop*** table contains levels of any other labels with Variable Type ‘Germplasm Descriptor’ for example other names or whether a certain entry is a control or check in a field trial, or its seed source (where you got the seed) or a label for the pedigree.

TRIAL DESIGN

The Trial Design component of the Logical Data Model defines the observation units in each dataset of the study and the experimental design and sample plans for those observation units. The information stored in the ***nd\_experiment*** and ***nd\_experiemntprop*** tables. General information about the study which is contained in study labels (variables with type Study Detail) is inherited by all the observation units of all datasets belonging to the study. Every dataset in a study has at least one observation unit (***nd\_experiement*** record). For field trials the Study Dataset has one observation unit, the Environment Dataset has one observation unit for each trial instance and the Plot Dataset has one observation unit for each plot.

Each record in the ***nd\_experiment*** table corresponds to one observation unit in a dataset and serves to link a specific combination of label values (levels) to a specific set of variate values. The levels of the design labels (treatments, design and lay-out) specify the context of the experiment or observation unit on which the associated values of the variates were observed. The study labels, environment labels and entry labels are associated by links from those components to the ***nd\_experiment*** records (observation units).

Figure 10. Observation units of the study, environment and plot datasets for study S9801.

| **nd\_experiment** | | |
| --- | --- | --- |
| **nd\_experiment\_id** | **nd\_geolocation\_id** | **type\_id** |
| 500 | 1 | 1010 |
| 501 | 1 | 1020 |
| 502 | 1 | 1155 |
| 503 | 1 | 1155 |
| 504 | 1 | 1155 |
| 505 | 1 | 1155 |
| 506 | 1 | 1155 |
| 507 | 1 | 1155 |
| 508 | 1 | 1155 |
| 509 | 1 | 1155 |
| 510 | 1 | 1155 |
| 511 | 1 | 1155 |
| 512 | 1 | 1155 |
| 513 | 1 | 1155 |

Each observation unit has a type indicated by ***nd\_experiment.type\_id*** linking to a ***cvterm***. Some types are:

1. Type = Study (***cvterm\_id***=1010) linking levels of labels which apply to the whole study (like the PI’s name) to observations which might have been made at the whole study level – like water condition eg irrigated or not.
2. Type = Dataset (***cvterm\_id***=1015) linking levels of the whole dataset environment to observations made at this level.
3. Type = Trial Environment (***cvterm\_id***=1020) linking levels of labels which apply to each instance of a trial (ie each environment where a trail is repeated). Usage of properties is similar to study above.
4. Type =Field Plot (***cvterm\_id***=1155). A field trial. Linked by proj\_id (to what?) and links into the stock table. The properties describe the plot. Raw observations are linked from here in the phenotype table.
5. Type=Sample. (***cvterm\_id***=1160) For a sample unit smaller than a plot
6. Type = Average (***cvterm\_id***=1170). 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.
7. Type = Summary (***cvterm\_id***=1180) for summary statistics like SEs and LSDs. Linked by proj\_id, links to the stock table, properties similar to plot.

Each observation unit (***nd\_experiment*** record) belongs to a project (study or dataset) and this is recorded in the ***nd\_experiment\_project*** table which allows many to many linkages between ***nd\_experiment*** records and projects although in the breeding context each observation unit will belong to only one dataset

Each ***nd\_experiment*** record links to Trial Environment and Location labels via the ***nd\_experiment.nd\_geolocation\_id*** field. If the location/environment information is not available or is not relevant (eg for the mean over several locations) then this field is set to 1 – a ‘not specified’ environment (since it is not allowed to be *null)*.

Details of the germplasm applied to each ***nd\_experiment*** record are linked via the ***nd\_experiment\_stock*** linkage table which allows many to many linkages between ***nd\_experiment*** records and stock records although in the breeding context this will almost always be one stock to many observation units (***nd\_experiment*** records). This table requires a type\_id for each link. It is not clear what information that should carry so we can set it to 1000 for now. (Note this can be used to solve the problem of identifying which stock belongs to which variable (stock label) in cases where there are multiple stocks related to the same nd\_experiemnet. Eg in a trial we may have a mixture of entries applied to each plot.)

Levels of other labels describing the experimental design are supplied in the ***nd\_experimentprop*** table and these are linked to their standard variable ID via the ***nd\_experimentprop.type\_id***. The ***nd\_experimentprop.rank*** is only needed if you have two values of the same ***type\_id*** for the same ***nd\_experiment\_id*** so it can be left 0 for the moment.

We make a distinction here between trial management conditions (eg fertilizer applied) which are uniform for a complete trial instance and which are recorded through labels with variable type ‘Environment Detail’ and those which are treatments in the experiment and so vary from one observation unit to another which are recorded through labels with variable type ‘Treatment Factor’. The former variables are stored in the ***nd\_geolocationprop*** table and the latter are part of the design and are stored in the ***nd\_experimentprop*** table as described above.

Figure 11. Design and treatment levels for plot units in the ***nd\_experimentprop*** table.

| **nd\_experimentprop** | | | | |
| --- | --- | --- | --- | --- |
| **nd\_experimentprop\_id** | **nd\_experiment\_id** | **type\_id** | **value** | **rank** |
| 2 | 502 | 8200 | 1 | 7 |
| 1 | 502 | 8210 | 1 | 5 |
| 3 | 502 | 8241 | 1 | 8 |
| 4 | 502 | 8260 | 100 | 9 |
| 6 | 502 | 61668 | 1 | 13 |
| 5 | 502 | 61670 | 1 | 10 |
| 8 | 503 | 8200 | 2 | 7 |
| 7 | 503 | 8210 | 1 | 5 |
| 9 | 503 | 8241 | 2 | 8 |
| 10 | 503 | 8260 | 200 | 9 |
| 12 | 503 | 61668 | 1 | 13 |
| 11 | 503 | 61670 | 2 | 10 |
| 14 | 504 | 8200 | 3 | 7 |
| 13 | 504 | 8210 | 1 | 5 |
| 15 | 504 | 8241 | 2 | 8 |
| 16 | 504 | 8260 | 200 | 9 |
| 18 | 504 | 61668 | 2 | 13 |
| 17 | 504 | 61670 | 1 | 10 |

...

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 64 | 512 | 8260 | 200 | 9 |
| 66 | 512 | 61668 | 3 | 13 |
| 65 | 512 | 61670 | 1 | 10 |
| 68 | 513 | 8200 | 12 | 7 |
| 67 | 513 | 8210 | 2 | 5 |
| 69 | 513 | 8241 | 1 | 8 |
| 70 | 513 | 8260 | 100 | 9 |
| 72 | 513 | 61668 | 3 | 13 |
| 71 | 513 | 61670 | 2 | 10 |

OBSERVATION

Variate values are recorded in the ***phenotype*** table. All phenotype records are linked to the appropriate ***nd\_experiment*** record via the ***nd\_experiment\_phenotype*** linkage table. This allows many to many linkages between ***nd\_experiment*** records and phenotype records although in breeding trials it will always be one ***nd\_experiment*** record to one or more phenotype records (possibly none). The type of the observation unit (***nd\_experiment*** record) indicates how the values are obtained. In the case of nd\_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).

The variate value is stored as a string in the ***phenotype.value*** field whether it is numeric or character except for categorical variates for which the valid values are stored in the ***cvterm*** table and the actual value is indicated by a link from ***phenotype.cvlaue\_id***.

The ***phenotype.observable\_id*** field links the phenotype value to the standard variable cvterm\_id in the ***cvterm*** table. The ***phenotype.name*** filed also stores the cvterm\_id for the standard variable but as a string. This enables easy linking the standard variable id in the ***projectprop.tyoe\_id field*** when browsing datasets.

Figure 12. Some phenotype values for plot data from study S9801 in the ***phenotype*** table.

| **phenotype** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **phenotype\_id** | **uniquename** | **Name** | **observable\_id** | **attr\_id** | **value** | **cvalue\_id** | **assay\_id** |
| 500 |  | 20370 | 20370 |  | 3 | 51211 |  |
| 501 |  | 20343 | 20343 |  | 80 |  |  |
| 502 |  | 18150 | 18150 |  | 10.3 |  |  |
| 503 |  | 20370 | 20370 |  | 3 | 51211 |  |
| 504 |  | 20343 | 20343 |  | 85 |  |  |
| 505 |  | 18150 | 18150 |  | 12.7 |  |  |
| 506 |  | 20370 | 20370 |  | 5 | 51213 |  |
| 507 |  | 20343 | 20343 |  | 103 |  |  |
| 508 |  | 18150 | 18150 |  | 18.7 |  |  |
| 509 |  | 20370 | 20370 |  | 4 | 51212 |  |
| 510 |  | 20343 | 20343 |  | 88 |  |  |
| 511 |  | 18150 | 18150 |  | 13.7 |  |  |
| 512 |  | 20370 | 20370 |  | 2 | 51210 |  |
| 513 |  | 20343 | 20343 |  | 79 |  |  |
| 514 |  | 18150 | 18150 |  | 12.6 |  |  |
| 515 |  | 20370 | 20370 |  | 1 | 51209 |  |
| 516 |  | 20343 | 20343 |  | 102 |  |  |
| 517 |  | 18150 | 18150 |  | 16.7 |  |  |
| 518 |  | 20370 | 20370 |  | 4 | 51212 |  |
| 519 |  | 20343 | 20343 |  | 87 |  |  |
| 520 |  | 18150 | 18150 |  | 19.2 |  |  |

OWNERSHIP AND ATTRIBUTION

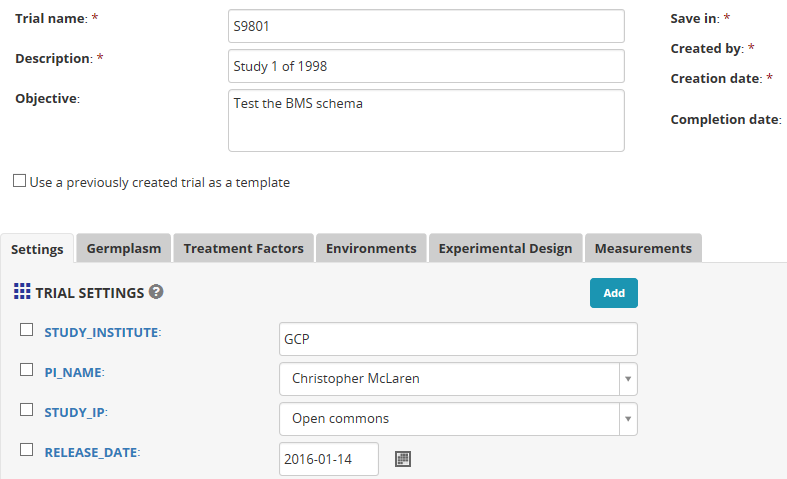
Every data value can be traced back to a database user. To do this, applications must add a UID variable at the appropriate level for the data value. Every study must have a Study\_UID (***cvterm\_id***=8020) variable, if a different user adds a dataset to the study, it must have a Dataset\_UID variable in the dataset with its value there also, if a different user adds a record to a dataset it must have an OU\_UID in the ***nd\_experimentprop*** table. If a different user adds a variable to a dataset is must have a Variable\_UID property attached to the variable description in the dataset and if a different user adds or changes a phenotype value .. we need a phenotypeprop table! Value\_UID

Any value then ‘belongs’ to its closest UID value in the order phenotype, variable, observation unit, dataset and study.

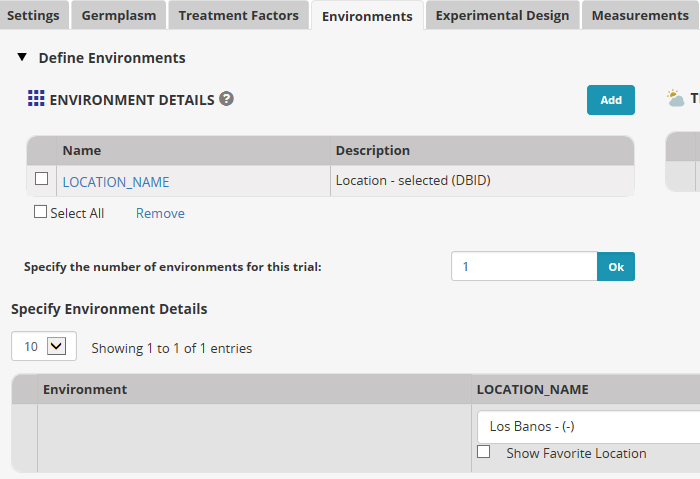
The same system could be used to timestamp data and different levels of precision.

**Annex 1: Tutorial to Load the S9801 Study**

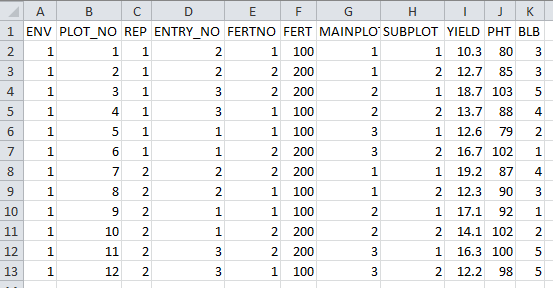
1. Use the Germplasm Import Toole to import germplasm A,B and C into a list called S9801.
2. Use the Ontology Manager Create two design variables MAINPLOT and SUBPLOT with property Blocking factor, method Assigned and new scales Main plot number and Sub-plot number respectively.
3. Add location “Los Banos” through Manage Program Settings.
4. Create a new trial called S9801 with title “Trial 1 of 1998’ and objective ‘Test the BMS schema’
5. Add Trial Setting Variables as shown in the figure below



1. Select the germplasm list S9801.
2. Add LOCATION\_NAME to the Environment details section of the Environment tab.



1. Prepare the following Lay-out file as a csv file:



1. Click ‘import’ on the Experimental Design tab and navigate to the above file, open it and then click continue.
2. Map ENV to TRIAL\_INSTANCE in the Environmental Factors group.
3. Map FERTNO to FERT\_NO and FERT to NFERT\_KG\_HA in the Design Factors group.
4. Map PHT to PH\_cm and BLB to BCLS\_1\_5 (I am using the maize database so you might have to use something else in another crop)
5. Other mappings have been automatically achieved and all look correct. Otherwise re-map them. While mapping you can change the local names of the variables to match the ones in Table 1 if desired.
6. Click Next. Review the measurements table and then click Finish.