Primary Functions supported by a Phenotyping Database (DMS)

1. Logistics and characterization data for Population Development

Breeding projects start with a number of crosses between strains of the crop which individually possess desirable traits for the target growing environment. The objective is to select offspring from these crosses which combine as far as possible the desirable traits from each parent. The logistics and data management for this process is facilitated by the Breeding Manager Application.

Usually the 'crossing table' is a simple data table indicating the identity of the male and female parents and some other details about the cross such as cross number, the date it was made, the number of seeds harvested, the name of the technician etc. If there is any Observation unit in this process it is the plot, pot or plant or plant part which bears the female flowers for the cross.

Population development proceeds through characterization and selection in a series of segregating populations or clonal evaluations depending on the inherent mating type of the crop. The database support is similar in all cases, so we can consider a pedigree breeding program where the segregating populations proceed from the direct progeny of the crosses – the F1s along a series of filial generations, F2, F3, F4 etc where each generation is produced by closed breeding within families of the previous generation so that the degree of homozygosity (inbreeding) increases rapidly through the populations.

Generally all the families of a particular generation (Fn) are treated together as a 'Nursery'. The nursery is planted out in rows with one family or line per 'row'. The Observation unit can therefore be taken as the 'row' which may literally be a row of plants, but could be a plot or pot also. There is minimal field design associated with nurseries, usually only a rectangular array of 'rows' interspersed at regular intervals with rows containing known check varieties for comparison with the test lines. The test lines are not usually randomized amongst the rows but rather appear in family groups – all lines from cross 1 followed by all lines from cross 2 etc. So the data table must store the identity of the row (observation unit), the identity of the line grown on the row, and whether it is a check variety or a test variety.

Generally nurseries are grown in a single location and single condition with a single replication of each test line since there is often not enough seed to repeat or replicate the nursery.

Some phenotypic traits may be recorded on nurseries, but these are usually traits which can be rapidly scored by eye such as phenology, plant architecture, disease susceptibility and so on. Often breeders' notes are recorded. All these variables may be stored, but the most important variables are logisitical – recording which rows (or plants within rows) are to be harvested and the seed kept for the next generation and which are to be discarded. At harvest, the amount of seed harvested from the retained rows or plants is often recorded.

The logistical data is queried for the purpose of identifying the next generation of lines (and the pedigree from the new line to its source line is recorded in the Genealogy Management System). The seed harvested data is queried to update the Inventory Management System, but the most important DMS query is a query back to previous generations (along the pedigree line recorded in the GMS) to see characterization data from source lines. For example a breeder will want to ask if a previous source line for the line he is currently considering showed resistance to a certain disease.

2. Evaluation Data for advanced lines

As the generations progress through the nurseries, fewer and fewer lines are retained and larger quantities of seed become available. At some stage field trial evaluations start to be conducted. The logistics and data management for Field trials is facilitated by the Fieldbook application.

The Fieldbook application merges a list of design and trait variables of interest to the breeder (the fieldbook template), with a list of advanced lines to be tested and possibly one or more check lines (the germplasm list) and a lay-out design – the arrangement of field plots into replicates, blocks, rows and columns as well as the random assignment of germplasm entries to those plots.

Field trials are often repeated in different environments to ensure a good sampling of the targeted growing environments. Each 'Trial Instance' tests the same germplasm, often with the same design but usually with a different randomization. The Fieldbook data table therefore consists of identification of the trial instance, the plot (observation unit) within each trial instance, the design details, the identity of the germplasm entry assigned to each plot, and then columns for the different traits to be measured on each plot. There may also be some logistical data such as breeders' notes, the identification of plots to be retained or discarded and the harvest of plots or plants from each plot.

Fieldbooks are generally exported and printed by Trial Instance, and shipped with the packed seed to the location where the trial will be grown. The data are often returned as an electronic spreadsheet for the individual instance.

One complication for the management of raw evaluation data is that the whole field plot is sometimes not the basic observation unit. Some traits are measured on subsamples of the plot such as randomly selected plants, or quadrats. It may even be that the number of these subsamples varies from plot to plot, and the number and type of subsamples may vary from trait to trait.

The most common query for raw evaluation data is to return subsets of the data from a single Field Trial for quality assurance, data transformation or single and multi-site analysis.

3. Derived and analysed data

Once the evaluation data have been collected, subjected to some local quality assurance, and saved into the DMS, it will be subjected to further analysis before decisions are made regarding the selection of the best performing lines. As mentioned above data sets are retrieved from the database and subjected to further quality assurance, and possibly data transformation. This latter function often combines values recorded on sampling units into 'averages' at the field plot level, and then derives new variables with different units or scales (conversion of g/plot to t/ha for example). Also individual raw variables are often combined for example ears per plant is calculated from the number of ears per plot and the number of plants per plot. The derived data must be stored back into the DMS alongside the raw data. This process is the first step in the analytical pipeline supported by the Breeding View Application.

The second step in the analysis pipeline supported by the breeding view Application is single site analysis. To support this, the DMS must be queried for subsets of data from a trial which contains the design variables, the treatment variables and the trait variable (raw or derived) to be analysed. Often you want to analyse a series of trial instances, so the query often needs to return a table indexed by trial instance (site/location/environment). The analysis produces tables of adjusted means indexed by treatment variables but averaged and adjusted for the design variables. The analyses also produce summary statistics such as standard errors of the means, heritabilities, and correlations between entries or between environments. The adjusted means and the summary statistics need to be stored back in the DMS. The results and statistics must be queried and viewed by the table Viewer Application to facilitate decision making.

The next steps in the analytical pipeline all involve querying the adjusted means and summary statistics for further analysis such as multi-site analysis or QTL analysis. These analyses produce more estimates and statistics which must be stored in the DMS or GDMS depending on the particular analysis.

4. Cross-study breeders' queries

All the functionalities described so far involve within-study queries of the DMS. However an important function of the DMS is to support some specific cross-study queries as well as general data mining applications. After some consultation we have come up with the following main queries required by many plant breeders. The phenotyping database will of course integrate with germplasm, genotype and location databases and the following queries will have to mesh with queries on those systems, but the current questions relate only to the phenotyping data. The following use cases are also rarely independent but are often linked so the query tool must be able to combine criteria from any combination of them.

- Find germplasm suitable for a geographical location i.e. certain environmental conditions. (Filter locations by the required environmental conditions, and find the germplasm with the best performance of some important trait(s) in those locations).

- Find germplasm with certain phenotypic characteristics i.e. traits that have been scored in trials. (Filter experimental observations by acceptable trait values and report the list of germplasm and associated trait values from those observations).
- Find available data for a given (list of) germplasm. (Filter experimental observations relating to the given germplasm and report values of traits observed).
- Head to head comparisons. (Find all experiments where a specific pair of lines has been compared for a certain trait and extract the set of pairwise comparisons with measures of statistical precision where possible).

5. General data integration

Finally, it will be desirable to have a free form query builder – The CropFinder Application which allows users to specify a set of reporting variables containing either environment, condition, design or trait values, and then filter the observations in the DMS according to constraints on multiple filter variables which can again cover environment, condition design or trait variables (which may be different from the reporting variables).

Integration of data in an abstract form involves retrieving and summarizing data values for a set of variates from a set of observation units in a coherent way so that the resulting data set can be interpreted in terms of properties of measured objects.

A system which supports data integration requires a data model, controlled annotation of model objects, a query tool and a summary mechanism.

The Data Model

The simplest, useful data model seems to be:

A Data Value is a measurement of a Variate on an Observation Unit.

Controlled Annotation of the Model Objects

A Data Value should be annotated with the following IP and Quality information:

- Who observed it?
- When?
- Who owns it?
- Who can see/use it?
- Why was it collected?
- Is it within the normal range for such data?
- How accurate is it?
- How precise is it?
- Is it fit for purpose?

A Variate is defined by:

- The property (trait) being measured
- The method of measurement
- The scale or unit of reporting

An Observation Unit is defined by a set of factor levels specifying the object being measured and the context in which it is measured. Factors are defined by:

- The property represented by their levels
- The method by which levels are assigned
- The scale or unit in which the levels are expressed

The Data Query

The Data Query problem is to retrieve data values for a set of report variates and levels of a set of report factors where the variates are measured on a set of observation units defined by one or more of the factors.

The set of report variates for which data should be retrieved may be specified more or less precisely by specifying a set of properties of interest with or without associated methods and scales and indeed could be determined by all variates which have data values on any or all of the specified observation units. The set of report factors for which levels should be retrieved can be similarly specified in terms of the properties, methods and scales of the factors or could be determined by all factors which have levels for any or all of the specified observation units.

The set of observation units could be defined as the set of observation units which have measurements for any or all of the specified report variates, or levels for any or all of the specified report factors or they could be defined by filtering on levels of a set of factors and/or on values of a set of variates (which may or may not overlap with the reporting factors and variates).

Data Summary

The Data Query produces a sparse matrix of data values from a list of annotated observation units (rows) for a set of result variates (columns). The row annotation consists of level values for a set of report factors. The real data integration problem is to summarize this matrix by merging columns and rows and averaging (in some way) multiple data values occurring in merged cells.

Columns which have the same property and same scale but different methods can, in principle be merged and continuous numeric values could be averaged, discrete or categorical values are another problem when values for the same observation unit differ. When the property is the same but the scales differ, then transformation to a common scale is an option. However the merging of columns is not the greatest problem, if it is done before rows, because there will seldom be multiple data values for the same observation unit.

Rows with common level annotations can, in principle be merged, but the common problem is to define a set of factors for which distinct level values must be kept apart, say the dependent factors (eg

germplasm ID) and a set over which values can be averaged, say the design factors (eg study ID). The second problem is what to do with rows for which some dependent factor levels are missing, and the third problem is knowing whether it makes sense to merge rows and average values across certain design variables (eg studies in well watered and drought prone environments). This latter problem is a problem for the biologist and the best the informaticist can do is present a tool which facilitates the decision.