Storage, analysis and communication of information from diverse wheat field trials

Micallef S 1,2, DeLacy I 1,2, Dieters M 1

1 School of Land, Crop and Food Sciences, The University of Queensland, Australia 2 Australian Centre for Plant Functional Genomics (ACPFG), The University of Queensland, Australia.

ABSTRACT

A procedure for the storage, analysis and communication of results from data generated from diverse field trials without compromising institutional confidentiality is discussed. Australian wheat breeders have been growing CIMMYT lines in their own trials for the past 40 or so years, yet only a small proportion of the trial data has made its way back to CIMMYT or other appropriate Australian research entities for further analysis. This information is important in determining which lines from CIMMYT are most suitable for the Australian environment. In addition, timely feedback from the Australian breeders will be useful for the CIMMYT breeders to make decisions on parents. The field trial data (raw data) or some kind of analysis from the raw data is what is normally distributed for further analysis, making it quite difficult to link this information to other research for comparison. A method which allows breeders to share their data without compromising their own institution’s privacy rules and which also allows linking of their results with other breeders’ results has been developed. All the information gathered on a group of wheat varieties is called a ‘study’. Each ‘study’ is divided into different data subsets, namely: (a) raw data, (b) intermediate data (entry BLUEs and weights from individual trials), (c) derived data and (d) environment data, but it is not necessary for each study to have all these datasets. In addition the availability of the intermediate data enables the analysis of subsets of the data from within or across studies without raw data. Examples of these across-study analyses are given. We also show how this information is collated together and can be queried to evaluate performance over a number of years and/or locations.

CIMMYT GERMLASM IN AUSTRALIA

The International Maize and Wheat Improvement Centre (CIMMYT) in Mexico has been working for 42 years with research institutions worldwide to improve the productivity and sustainability of wheat breeding systems, through a very successful global research program. Through a partnership with CIMMYT and the Grains Research and Development Corporation (GRDC), Australia conducts field trials as part of its scientific research program to identify wheat varieties adapted to the diverse Australian regional environments. The germplasm received from CIMMYT, comes from a variety of different trials, namely: (a) the Elite Spring Wheat Yield Trials (ESWYT), (b) Semi Arid Wheat Yield Trails (SAWYT), (c) High Temperature Wheat Yield Trials (HTWYT), (d) High Rainfall Wheat Yield Trial (HRWYT) and (e) International Bread Wheat Screening Nursery (IBWSN).

The GRDC has a number of research projects involving CIMMYT material. These projects aim to evaluate the imported CIMMYT germplasm for Australian environments, improve the quality of wheat, and enhance stress and disease tolerance. With the involvement of many research institutions around Australia, GRDC responded to the need of good communication between all participating partners. The information collected from the trials around Australia is now collected and made available publicly via the internet. The webpage developed is called CAGE (CIMMYT and Australia Germplasm Enhancement) and can be found at http://cage.lafs.uq.edu.au.

This information is also returned to CIMMYT and used by the breeders to help make selection and crossing decisions for the next breeding cycle.

This system promotes collaboration amongst Australian breeders and scientists working with CIMMYT material, ensures access to superior germplasm which allows them to be competitive in global markets, and encourages this germplasm to be included in their own trials, as well as reselection and crossing with their own material.

DATA COLLECTION

The International Wheat Information System (IWIS) used at CIMMYT manages and integrates information about all CIMMYT lines. Each line is given a unique identifier which enables breeders and data users to pinpoint a particular line even though it might have several names.

The germplasm imported into Australia from CIMMYT is sent directly to quarantine, where a quarantine code (QCode and QNo) is issued. This quarantine code is then linked back to the unique identifier (GID) issued by CIMMYT, providing traceability throughout the importation and distribution process.

The imported germplasm list (fig1) is distributed to the project partners together with a template for the collection of field data. This template file is set up as follows:

- an observation sheet (fig 2) where all data is recorded; and
- a description sheet (fig 3) which describes the data in each column of the observation sheet.

Each trait being recorded in the observation sheet (e.g. Yield_BLUE) needs to be linked to a trait name from an
ontology list (e.g. GRAIN_YIELD), and also have a scale (e.g. t/ha or grams/plot) and method (e.g. BLUE or BLUP) defined.

DATA STORAGE

All the trials of one project are stored in the system as one ‘study’1. A study is a collection of one or more datasets collected from scientific experiments, field trials, environment data (such as rainfall statistics), laboratory testing and data analysis. The division of data into different sets allows the users to conveniently manage the data received at different times and from different places. In this project, data from three GRDC-CIMMYT projects were collected and entered in the database as three studies – the International Adaptation Trials (IAT), the Germplasm Enhancement Trials (GET) and the Synthetics Trials. Each project had a minimum of 3 datasets:

- raw data: the data collected by the breeders from the trials with row/column/plot values;
- intermediate data: grain yield BLUE (Best Linear Unbiased Estimator) values, number of observations per line in the trial and the residual variance of the site or trial being considered (known as triplets, Fig 3); and
- Environment data: information about the trial site such as GPS information, rainfall measurements, planting and harvest dates, plot area, trial design, fertilizers and irrigation applied and information on plant damage.

The derived data or complete analysis of each project can be stored as another dataset in each respective study. If the derived data consists of meta-analysis (i.e. combined analysis of data from a number of studies), then the resulting derived dataset can be stored in a separate study.

DATA ANALYSIS – INTERMEDIATE DATA

A two-stage analysis (Smith et al 2005) is often more convenient or even necessary for the analysis of large plant breeding field trial data sets when conducting a

---

1. Study
mixed model analysis using REML. In this analysis the first stage consists of conducting a search for a preferred model (Qiao et al 2000) of each trial using a REML analysis with entries coded as fixed effects. To complete the analysis a second stage is conducted using a set of triplets from each trial to conduct a weighted analysis of the entries across all trials. The set of triplets from each trial consists of the BLUE $b_{ij}$ for the $i^{th}$ entry in the $j^{th}$ trial, the number of replicates $r_{ij}$ from which each $b_{ij}$ was estimated and the residual variance $v_{j}$ for that trial. The $r_{ij}$ and $v_{j}$ are used to calculate the weights for the second stage. If the set of triplets from the first stage analysis of all trials are stored as data sets for all studies it enables researches to conduct a second stage analysis of any set of entries for any set of trials (say a region) for any set of years by simply downloading all the triplets pertaining to the entries in the region of study for the time period of interest. Another major advantage provided by calculating the set of triplets results from the ability to store only the data from a trial for the subset of entries that are to be made public. This enables participating research and breeding programs to enter the CIMMYT introduction lines in their normal trials and to return only the data pertaining to these lines. This eliminates the need to grow ‘public’ lines in special trials.

COMMUNICATION

All the information collected from the GET, IAT and Synthetics projects, has been organised in a website – http://cage.lafs.uq.edu.au. (fig4) The website is hosted at the University of Queensland and is continuously being updated as soon as data become available.

The CAGE website contains information about the GRDC-CIMMYT projects, lists of the material received from CIMMYT and their assigned quarantine codes, CIMMYT breeders data, field data, intermediate data and environment data from Australian trials, information on upcoming project meetings and all presentations from past meetings, Diversity Arrays Technology (DArT) data results for CIMMYT material and a link to the GWIS database for searching pedigree information about any wheat variety.

REFERENCES


Fig 4: CAGE website at http://cage.lafs.uq.edu.au