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 GERMPLASM 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

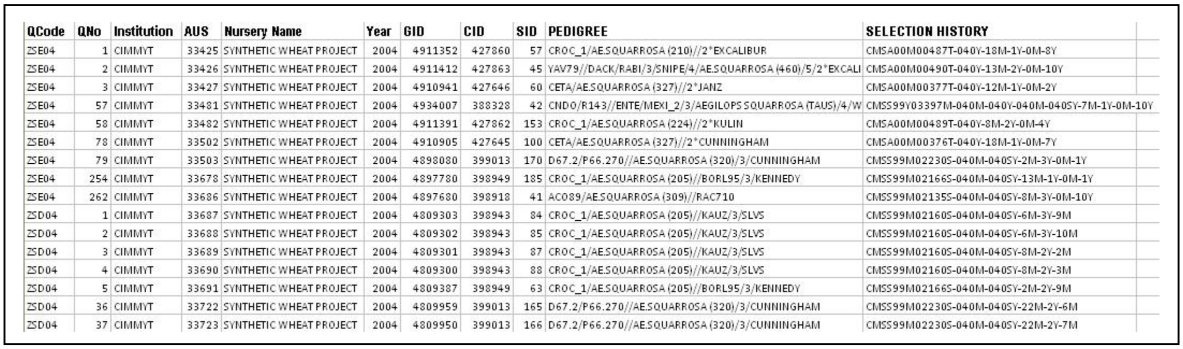
The International Wheat Information System (IWIS) used at CIMMYT manages and integrates information about all CIMMYT lines. Each line is given a unique identifier which enables breeders and data users to pinpoint a particular line even though it might have several names.

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:

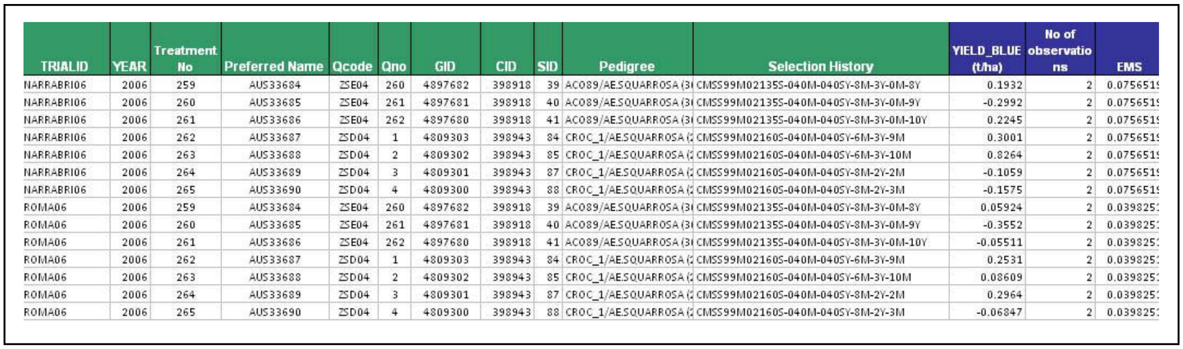
- an observation sheet (fig 2) where all data is recorded; and
- a description sheet (fig 3) which describes the data in each column of the observation sheet.

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.



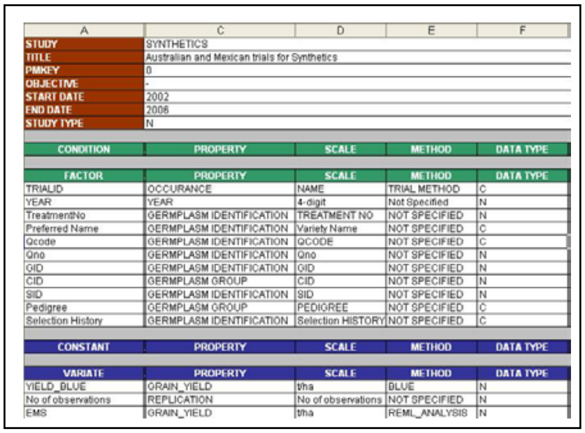
QCode	QNo	Institution	AUS	Nursery Name	Year	GID	CID	SID	PEDIGREE	SELECTION HISTORY
ZSE04	1	CIMMYT	33425	SYNTHETIC WHEAT PROJECT	2004	4911352	427860	57	CROC_1/AE.SQUARROSA (210)//2*EXCALIBUR	CMSA00M00487T-040Y-18M-1Y-0M-8Y
ZSE04	2	CIMMYT	33426	SYNTHETIC WHEAT PROJECT	2004	4911412	427863	45	YAV79//DACK/RABI/3/SNIPE/4/AE.SQUARROSA (460)/5/2*EXCALI	CMSA00M00490T-040Y-13M-2Y-0M-10Y
ZSE04	3	CIMMYT	33427	SYNTHETIC WHEAT PROJECT	2004	4910941	427646	60	CETA/AE.SQUARROSA (327)//2*JANZ	CMSA00M00377T-040Y-12M-1Y-0M-2Y
ZSE04	57	CIMMYT	33481	SYNTHETIC WHEAT PROJECT	2004	4934007	388328	42	CNDO/R143//ENTE/MEXI_2/3/AEGILOPS SQUARROSA (TAUS)/4/W	CMSS99Y03397M-040M-040Y-040M-040SY-7M-1Y-0M-10Y
ZSE04	58	CIMMYT	33482	SYNTHETIC WHEAT PROJECT	2004	4911391	427862	153	CROC_1/AE.SQUARROSA (224)//2*KULIN	CMSA00M00489T-040Y-8M-2Y-0M-4Y
ZSE04	78	CIMMYT	33502	SYNTHETIC WHEAT PROJECT	2004	4910905	427645	100	CETA/AE.SQUARROSA (327)//2*CUNNINGHAM	CMSA00M00376T-040Y-18M-1Y-0M-7Y
ZSE04	79	CIMMYT	33503	SYNTHETIC WHEAT PROJECT	2004	4898080	399013	170	D67.2/P66.270//AE.SQUARROSA (320)/3/CUNNINGHAM	CMSS99M02230S-040M-040SY-2M-3Y-0M-1Y
ZSE04	254	CIMMYT	33678	SYNTHETIC WHEAT PROJECT	2004	4897780	398949	185	CROC_1/AE.SQUARROSA (205)//BORL95/3/KENNEDY	CMSS99M02166S-040M-040SY-13M-1Y-0M-1Y
ZSE04	262	CIMMYT	33686	SYNTHETIC WHEAT PROJECT	2004	4897680	398918	41	AC089/AE.SQUARROSA (309)//RAC710	CMSS99M02135S-040M-040SY-8M-3Y-0M-10Y
ZSD04	1	CIMMYT	33687	SYNTHETIC WHEAT PROJECT	2004	4809303	398943	84	CROC_1/AE.SQUARROSA (205)//KAUZ/3/SLVS	CMSS99M02160S-040M-040SY-6M-3Y-9M
ZSD04	2	CIMMYT	33688	SYNTHETIC WHEAT PROJECT	2004	4809302	398943	85	CROC_1/AE.SQUARROSA (205)//KAUZ/3/SLVS	CMSS99M02160S-040M-040SY-6M-3Y-10M
ZSD04	3	CIMMYT	33689	SYNTHETIC WHEAT PROJECT	2004	4809301	398943	87	CROC_1/AE.SQUARROSA (205)//KAUZ/3/SLVS	CMSS99M02160S-040M-040SY-8M-2Y-2M
ZSD04	4	CIMMYT	33690	SYNTHETIC WHEAT PROJECT	2004	4809300	398943	88	CROC_1/AE.SQUARROSA (205)//KAUZ/3/SLVS	CMSS99M02160S-040M-040SY-8M-2Y-3M
ZSD04	5	CIMMYT	33691	SYNTHETIC WHEAT PROJECT	2004	4809387	398949	63	CROC_1/AE.SQUARROSA (205)//BORL95/3/KENNEDY	CMSS99M02166S-040M-040SY-2M-2Y-9M
ZSD04	36	CIMMYT	33722	SYNTHETIC WHEAT PROJECT	2004	4809959	399013	165	D67.2/P66.270//AE.SQUARROSA (320)/3/CUNNINGHAM	CMSS99M02230S-040M-040SY-22M-2Y-6M
ZSD04	37	CIMMYT	33723	SYNTHETIC WHEAT PROJECT	2004	4809950	399013	166	D67.2/P66.270//AE.SQUARROSA (320)/3/CUNNINGHAM	CMSS99M02230S-040M-040SY-22M-2Y-7M

Fig 1: Example of Quarantine list



TRIALID	YEAR	Treatment No	Preferred Name	Qcode	Qno	GID	CID	SID	Pedigree	Selection History	YIELD_BLUE (t/ha)	No of observations	EMS
NARRABRI06	2006	259	AUS33684	ZSE04	260	4897682	398918	39	AC089/AE.SQUARROSA (3	CMSS99M02135S-040M-040SY-8M-3Y-0M-8Y	0.1932	2	0.075651
NARRABRI06	2006	260	AUS33685	ZSE04	261	4897681	398918	40	AC089/AE.SQUARROSA (3	CMSS99M02135S-040M-040SY-8M-3Y-0M-9Y	-0.2992	2	0.075651
NARRABRI06	2006	261	AUS33686	ZSE04	262	4897680	398918	41	AC089/AE.SQUARROSA (3	CMSS99M02135S-040M-040SY-8M-3Y-0M-10Y	0.2245	2	0.075651
NARRABRI06	2006	262	AUS33687	ZSD04	1	4809303	398943	84	CROC_1/AE.SQUARROSA (2	CMSS99M02160S-040M-040SY-6M-3Y-9M	0.3001	2	0.075651
NARRABRI06	2006	263	AUS33688	ZSD04	2	4809302	398943	85	CROC_1/AE.SQUARROSA (2	CMSS99M02160S-040M-040SY-6M-3Y-10M	0.8264	2	0.075651
NARRABRI06	2006	264	AUS33689	ZSD04	3	4809301	398943	87	CROC_1/AE.SQUARROSA (2	CMSS99M02160S-040M-040SY-8M-2Y-2M	-0.1059	2	0.075651
NARRABRI06	2006	265	AUS33690	ZSD04	4	4809300	398943	88	CROC_1/AE.SQUARROSA (2	CMSS99M02160S-040M-040SY-8M-2Y-3M	-0.1575	2	0.075651
ROMA06	2006	259	AUS33684	ZSE04	260	4897682	398918	39	AC089/AE.SQUARROSA (3	CMSS99M02135S-040M-040SY-8M-3Y-0M-8Y	0.05924	2	0.039825
ROMA06	2006	260	AUS33685	ZSE04	261	4897681	398918	40	AC089/AE.SQUARROSA (3	CMSS99M02135S-040M-040SY-8M-3Y-0M-9Y	-0.3552	2	0.039825
ROMA06	2006	261	AUS33686	ZSE04	262	4897680	398918	41	AC089/AE.SQUARROSA (3	CMSS99M02135S-040M-040SY-8M-3Y-0M-10Y	-0.05511	2	0.039825
ROMA06	2006	262	AUS33687	ZSD04	1	4809303	398943	84	CROC_1/AE.SQUARROSA (2	CMSS99M02160S-040M-040SY-6M-3Y-9M	0.2531	2	0.039825
ROMA06	2006	263	AUS33688	ZSD04	2	4809302	398943	85	CROC_1/AE.SQUARROSA (2	CMSS99M02160S-040M-040SY-6M-3Y-10M	0.08609	2	0.039825
ROMA06	2006	264	AUS33689	ZSD04	3	4809301	398943	87	CROC_1/AE.SQUARROSA (2	CMSS99M02160S-040M-040SY-8M-2Y-2M	0.2964	2	0.039825
ROMA06	2006	265	AUS33690	ZSD04	4	4809300	398943	88	CROC_1/AE.SQUARROSA (2	CMSS99M02160S-040M-040SY-8M-2Y-3M	-0.06847	2	0.039825

Fig 2: Example of a dataset with intermediate data, showing the triplet: Yield BLUE, number of observations and residual variance.



A	C	D	E	F
STUDY	SYNTHETICS			
TITLE	Australian and Mexican trials for Synthetics			
PMKEY	0			
OBJECTIVE	-			
START DATE	2002			
END DATE	2006			
STUDY TYPE	N			
CONDITION	PROPERTY	SCALE	METHOD	DATA TYPE
FACTOR	PROPERTY	SCALE	METHOD	DATA TYPE
TRIALID	OCCURANCE	NAME	TRIAL METHOD	C
YEAR	YEAR	4-digit	Not Specified	N
TreatmentNo	GERMPLASM IDENTIFICATION	TREATMENT NO	NOT SPECIFIED	N
Preferred Name	GERMPLASM IDENTIFICATION	Variety Name	NOT SPECIFIED	C
Qcode	GERMPLASM IDENTIFICATION	QCODE	NOT SPECIFIED	C
Qno	GERMPLASM IDENTIFICATION	Qno	NOT SPECIFIED	N
GID	GERMPLASM IDENTIFICATION	GID	NOT SPECIFIED	N
CID	GERMPLASM GROUP	CID	NOT SPECIFIED	N
SID	GERMPLASM IDENTIFICATION	SID	NOT SPECIFIED	N
Pedigree	GERMPLASM GROUP	PEDIGREE	NOT SPECIFIED	C
Selection History	GERMPLASM IDENTIFICATION	Selection HISTORY	NOT SPECIFIED	C
CONSTANT	PROPERTY	SCALE	METHOD	DATA TYPE
VARIATE	PROPERTY	SCALE	METHOD	DATA TYPE
YIELD_BLUE	GRAIN_YIELD	t/ha	BLUE	N
No of observations	REPLICATION	No of observations	NOT SPECIFIED	N
EMS	GRAIN_YIELD	t/ha	REML_ANALYSIS	N

Fig 3: Description sheet of data collection template file

DATA STORAGE

All the trials of one project are stored in the system as one ‘study’. 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 3 different 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 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 seaching pedigree information about any wheat variety.

REFERENCES

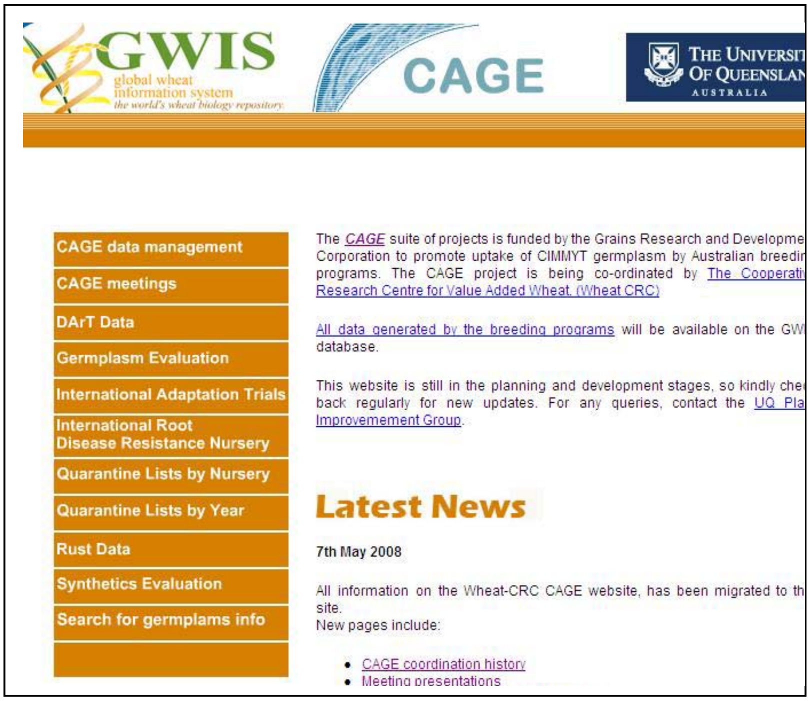
Bruskiewich RM, Cosico AB*, et al.* (2003) Linking genotype to phenotype: the International Rice Information System (IRIS). *Bioinformatics* **19 Suppl 1**, i63-5.

CIMMYT. 2001. Research Highlights of the CIMMYT Wheat Program, 1999-2000. Mexico, D.F.

McLaren CG, Bruskiewich RM, Portugal AM, Cosico AB (2005) The International Rice Information System. A platform for meta-analysis of rice crop data. *Plant Physiol* **139**, 637-42

Qiao CG, Basford KE, DeLacy IH (2000) Evaluation of experimental designs and spatial analyses in wheat breeding trials. *Theoretical and Applied Genetics* 100: 9-16.

Smith AB, Cullis BR, Thompson R. (2005) The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *J Agr Sci* 143: 449-462.



The CAGE suite of projects is funded by the Grains Research and Developme Corporation to promote uptake of CIMMYT germplasm by Australian breedir programs. The CAGE project is being co-ordinated by The Cooperati Research Centre for Value Added Wheat. (Wheat CRC)

All data generated by the breeding programs will be available on the GW database.

This website is still in the planning and development stages, so kindly che back regularly for new updates. For any queries, contact the UQ Pla Improvement Group.

Latest News

7th May 2008

All information on the Wheat-CRC CAGE website, has been migrated to th site.
New pages include:

- CAGE coordination history
- Meeting presentations

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