The International Crop Information System manages genealogical, phenotypic and genotypic data in a wheat breeding program

Clarke FR[1], Yates SM[1], Clarke JM[1], Knox RM, DePauw RM[1], and McLaren CG[2]

[1]Agriculture and Agri-Food Canada, Semiarid Prairie Agricultural Research Centre, P.O. Box 1030, Swift Current, Saskatchewan, Canada S9H 3X2, [2]International Rice Research Institute, DAPO Box 7777, Metro Manila, Philippines

ABSTRACT

The International Crop Information System (ICIS) links pedigrees to phenotypic (agronomic, disease and end-use functionality) and genotypic data that is made easily accessible. The Genealogy Management System (GMS) interfaces 'Central' public and 'Local' private databases which facilitates the global sharing of non-sensitive pedigrees, selection histories and other descriptors in the Central database while interfacing with Local databases which contain the sensitive data. The Central GMS now contains more than 5.8 million hexaploid (*Triticum aestivum* L.) and durum (*T. turgidum* L. ssp. *durum (Desf.)Husn.*) wheat genotypes, and our local GMS has more than 30,000 entries. The Data Management System (DMS), which has public phenotypic data for 491 nurseries spanning 1969 to 2006, manages the phenotypic and genotypic data and links it to lines in the GMS. The Data Comparison Tool (5.5) provides three queries to either compare two genotypes, to retrieve data by trial, or to retrieve data for a list of genotypes, and outputs the phenotypic and genotypic information to Microsoft Excel or other formats such as text. We routinely apply ICIS to aid in choice of parents for crossing and for managing and linking the genotypic data from haplotyping, association mapping and marker-development projects.

INTRODUCTION

Wheat breeding programs generate massive amounts of data and information, ranging from pedigree ancestry to phenotypic and genetic data, which is used in cultivar improvement. With the increased use of computers in data collection and analysis, an electronic information system is a practical and efficient option for storage and summarization of the data.

From 2000 to 2003, the International Crop Information System (ICIS) was evaluated at the Semi-arid Prairie Agricultural Research Centre (SPARC), in Swift Current, Saskatchewan. ICIS is an open-source breeding information system developed primarily by a team at the International Rice Research Institute (IRRI) in the Philippines (McLaren et al 2005). During this assessment, the value of having an information system such as ICIS became evident and so in 2003, the decision was made to integrate its use into the SPARC durum and hexaploid wheat breeding programs.

In 2004, ICIS was expanded to include users at the Lethbridge Research Centre (LRC) in Lethbridge, Alberta and at the Cereal Research Centre (CRC) in Winnipeg, Manitoba, with both centres to the databases at SPARC. Since 2000, SPARC has been part of an international ICIS community of users who are routinely submitting feature requests, bug reports and discussions to the Crop Research Informatics Laboratory (CRIL), an alliance of the IRRI developers and a team at the International Maize and Wheat Improvement Centre (CIMMYT) in Mexico, who now oversee the development of ICIS.

THE SYSTEM

The ICIS implementation at SPARC is divided into three main sections; a Genealogy Management System (GMS), a Data Management System (DMS), and a Gene Management System (GEMS).

The function of the Genealogy Management System (GMS) is to store pedigree information, such as ancestry, breeding methods, date and origin of germplasm, as well as references and notes. At SPARC, this information is entered into the database primarily by using the ICIS applications BROWSE and SETGEN. SETGEN is a multi-function application that not only allows the user to enter lines individually or through a batch mode, but also allows the user to view or edit the information, create crossing blocks and advance derivative lines by generation. SETGEN also allows the user to create lists of cultivars, and displays them through a tree interface, similar to the Microsoft Windows GUI. BROWSE, a Fortran program, runs in a DOS window and is an alternative for entering and editing of pedigree information. BROWSE allows users to run other functions, such as Coefficient of Parentage, Mendelgrams, and it will also query the GMS for sister lines, for descendants or for a specific cross. The GMS is comprised of two databases: a Central GMS and a Local GMS. The Central GMS, recently named IWIS3, is distributed by CIMMYT, and contains information on thousands of wheat cultivars from around the world. This database is considered public information and is shared internationally with all ICIS wheat users. Local GMS databases, administered at SPARC, contain durum and hexaploid pedigrees from the breeding programs at

SPARC, LRC and CRC. Much of this information is considered sensitive and is only available to users within Agriculture and Agri-Food Canada (AAFC). The present rule for SPARC breeders is that any of their pedigree information that is at least five years old can be considered public and is sent to CIMMYT to be uploaded into future releases of IWIS3.

The Data Management System (DMS) is used for storage of almost any type of data related to the germplasm stored in the GMS. The SPARC DMS stores phenotypic data from field trials, as well as end-use quality and genetic data collected from the laboratories at SPARC, Grain Research Laboratory (GRL) and CRC. The DMS is comprised of two Central databases for all AAFC users and two Local databases for each research centre. Unlike IWIS3, all DMS databases are administered at SPARC, and consist of only AAFC data. This data is not shared internationally with other ICIS wheat users. The Central databases each contain genetic and phenotypic data, respectively, that the researchers wish to share with all AAFC users, for example cultivar registration trial data. The local databases, however, contain genetic and phenotypic data that is pertinent to users at a particular research station or group. Data from ongoing projects or from early generation field trials might be included in the local DMS databases. All phenotypic and genetic studies are loaded into the DMS using the DMS Workbook, a Microsoft Excel-based tool that will check the dataset for errors and highlight any problems for the user. The DMS Workbook will also retrieve datasets and allow the user to create templates for studies that have common variables.

The last piece of the SPARC ICIS implementation is the Gene Management System (GEMS). Marker-assisted selection has become an integral part of wheat breeding, and there has been a strong push to include such information in ICIS. The function of GEMS is to store molecular information related to the studies found in the Genetic DMS databases. Currently, GEMS stores information on the specific markers being used, as well as information about the molecular variants they produce. This data is loaded into the GEMS database from the DMS Workbook through special functions. New tools are being developed at CRIL and SPARC to expand the scope of this data. Also, there is potential to store information about the polymerase chain reaction (PCR) protocols being used in the laboratory (i.e. PCR Recipe, PCR conditions, Electrophoresis, Gel Recipe, Gel Photos). The GEMS has only a central database and is available to all AAFC users.

At SPARC, there is only one ICIS administrator that has write access to the databases. This is to ensure the reliability and integrity of the information stored within each database and to keep a standardized set of protocols in place for loading and editing the data across breeding programs. This is in contrast to the classical approach to ICIS, where each breeder is the administrator of their own data.

UTILIZATION

As with any meaningful information system, the first step was to collate historical data and organize it in a way that made sense to not just the breeders, but also to the technicians and other researchers who would be accessing it across AAFC research centres. This was no trivial task, and has become an on-going exercise as more information is discovered from various sources.

The next challenge was to enter thousands of pedigrees and supporting information from the hexaploid and durum breeding programs at SPARC into the local GMS. This included going through all electronic and paper copies available, with some going back as far as the early 1900s. Each time one of these cultivars was entered, ICIS concurrently searched the Central Wheat GMS database for any matches in either the cultivar name or its parents. In many cases, matches found in the Central GMS database linked to previously unknown information thereby allowing breeders to have a larger picture of the ancestry of the cultivars. With more pedigree data being shared among ICIS wheat collaborators, and regular distribution of updates, the Central Wheat GMS has become an invaluable tool for all wheat breeders.

Following entry of pedigree information into the local GMS, phenotypic data was entered into the DMS. With the DMS being very flexible and a massive amount of data to enter, it was extremely important to plan in advance which data was relevant and how to store it so that meaningful queries could be developed. One of the major hurdles was the inconsistent acronyms used over time by different breeding programs, labs and technicians. To help with standardization of these acronyms, DMS Workbook Templates were developed for both the hexaploid and durum programs, which include all phenotypic traits measured in a growing season. These templates are completed by the technicians that handle the phenotypic data. After the data is finalized, they use simple SAS (version 8.2, Littell et al. 1996) routines to export the information into the template in a matter of seconds. The ICIS administrator then checks the formatting of the template and then loads it into the DMS with the DMS Workbook.

Entering genetic data into ICIS has started on a very limited basis because GEMS is still being incorporated into the SPARC ICIS scheme. The GEMS database structure allows for much more genetic information to be loaded than can be managed with currently available software tools, so a collaborative effort is underway to create these tools. Templates have been developed at SPARC to load genetic studies into the DMS and GEMS databases, and the Data Comparison Tool is being modified to query and output this information. Currently, the genetic information at SPARC has been limited to marker evaluation studies of prospective

parents and haplotyping studies, although the potential is available for nearly all types of genetic information (SSR, AFLP, SNP, DaRT).

The Data Comparison Tool was first developed at SPARC in 2002 to query the DMS. The purpose of the application is to query the databases for phenotypic data and output the results in a simple summary spreadsheet. The primary goal was to create a user-friendly, intuitive query application that did not require the user to have extensive knowledge of the ICIS schemata. The application was originally created using MS SQL in Microsoft Access, then it was re-written into a Visual Basic for Applications (VBA) program in Microsoft Excel and finally it has been migrated to the VB.Net programming environment.

The Data Comparison Tool has four main queries:

- List Studies in ICIS - gives a full list of all phenotypic and genetic studies in the ICIS DMS databases.
- Line vs. Line Comparison - allows the user to compare phenotypic and genetic data of selected traits for cultivars over the same studies. It also allows the user to include the checks from each study, if available.
- Study Retriever - retrieves the phenotypic and genetic data of selected traits from all cultivars grown in a specified field trial (or study).
- Germplasm List Retriever - queries the databases for phenotypic and genetic data of selected traits for a list of germplasm (which is created in SETGEN). The user can also get an output of all studies in which the listed germplasm is present.

At SPARC, the Data Comparison Tool is used to compare potential parents when planning crosses. It is also used in preparation of historical summaries of data for new cultivars, eliminating the arduous task of searching through numerous electronic files and paper copies as was done previously. It has major applications for quantitative genetic and other studies using historical data. The tool is flexible enough to work for almost any data entered into the DMS, and can be modified to meet new and changing needs of researchers.

The growing global support network of ICIS users has become one of its greatest strengths. While CRIL handles majority of the development and technical support for ICIS, users are also able to interact and get support through two websites, CropForge and CropWiki, both run by CRIL.

CropWiki (http://cropwiki.irri.org), which has a format similar to WikiPedia (http://en.wikipedia.org), contains the ICIS Technical Documentation Manual (TDM), the minutes and presentations from workshops, recaps of discussions among users, as well as tutorials and future proposals for ICIS. As is the case with WikiPedia, CropWiki allows registered users to add and edit information on the site to help improve and expand the information available.

CropForge (http://cropforge.org) is a more technical site that brings the developers and users together for bug reporting, feature requests, support requests, sharing of code through CVS, and to have technical discussions in the forums. The latest releases of ICIS applications are available for download here. CropForge is divided into various project pages (the Data Comparison Tool being one), with the ICIS Communication page being the main page for interaction with developers.

The collaborative nature of the ICIS community has benefited all involved, from the beginning user to the more advanced, and has been essential to the expansion of ICIS at SPARC. Although the ICIS community is global, feedback is usually very prompt, which leaves very little down time while dealing with obstacles.

SUMMARY

As data generation balloons as breeding programs increase the collection of phenotypic and genotypic data, it is imperative that researchers find proficient ways to handle storage and summarization of data. The implementation of the International Crop Information System at SPARC has proven to be an efficient way of managing the massive volume of data generated by the breeding program and associated marker development projects. The response from other users at AAFC has also been positive because of the flexibility that ICIS provides to adapt to diverse needs of researchers.

While there is a learning curve associated with starting up an ICIS implementation, with support from CRIL, the international ICIS community, the CropForge and Crop Wiki sites, the annual workshops and the technical documents, new users can implement a functioning system in a relatively short period of time.

ACKNOWLEDGEMENTS

We gratefully acknowledge funding of this research by Agriculture and Agri-Food Canada, the Western Grains Research Foundation, and the Agriculture and Agri-Food Canada Matching Investment Initiative.

REFERENCES

McLaren, C.G., Bruskiewich, ,R.M., Portugal, A.M., Cosico, A.B. 2005. The International Rice Information System. A platform for meta-analysis of rice crop data. Plant Physiology 139:637-642.

SAS (r) 9.1 (TS1M3), Copyright (c) 2002-2003. SAS Institute Inc., Cary, NC, USA.