MANAGING VARIETY IMPROVEMENT PROGRAM OPERATIONS USING AN INTEGRATED DATABASE AND HANDHELD COMPUTERS TO IMPROVE EFFICIENCY AND REDUCE ERRORS

By

PHIL LETHBRIDGE[1], MIKE COX[2]
BSES Limited Meringa[1] and Bundaberg[2]
plethbridge@bses.org.au

KEYWORDS: Variety Improvement, Database, SPIDNet, SPIDNetMobile.

Abstract

SPIDNET IS a suite of software solutions created by BSES Limited for the management of sugarcane variety improvement data. It was programmed using Microsoft's .NET Framework, the C# programming language and SQL Server 2005 relational database. Development commenced in 2003 with web and desktop applications and has now extended to mobile devices running the Windows Mobile 5.0 operating system. This paper describes the main features of SPIDNet and recent developments using SPIDNetMobile and handheld computers. Basic functionalities of SPIDNet include: managing plantings, plots and observations; pedigree management; cross pollination; and traceability. Symbol MC70 handheld computers, equipped with barcode scanners and radios for Bluetooth and WiFi communication, have provided an increase in data collection efficiency and accuracy. Data loss is minimised using Zebra RW420 mobile printers equipped with Bluetooth radios to print valuable information as it is collected. SPIDNet allows users to define new plantings and upload field plans as .csv files for validation and storage in a centralised database. Microsoft SQL Server provides replication features to copy data from this central database to smaller Compact Edition databases on PDAs as required. Robust labels and bar-coded unique identifiers enable SPIDNetMobile to display associated data in the field, and enforce relational integrity as new data are collected. Data collection operations currently include automatic capture of plot weights at harvest, validated capture of source plot data as new plantings are planted, template-based data capture permitting capture of any custom data against a given field plan, parent census and pollen test results during cross pollination operations, and other small utilities.

Introduction

Variety improvement programs for all crops involve large numbers of genotypes, trials and trial plots, for which large volumes of information are collected. In addition, pedigree information is usually very important. Sugarcane breeding programs are no different and appropriate, secure systems are critical to manage these

large amounts of information. For example, in the four regional selection programs (Northern, Burdekin, Central and Southern), we annually grow approximately:

- 100 000 seedlings in 4500 family plots involving 1000 different crosses (Stage 1—Progeny Assessment Trials, PAT)
- 20 000 clonal single-row plots (Stage 2—Clonal Assessment Trials, CAT)
- 4000 clonal 4-row plots (Stage 3—Final Assessment Trials)
- 8000 propagation plots to supply planting materials for trials
- 1500 plots in arrowing blocks for crossing
- 4000 plots for disease screening.

As sugarcane is grown over ratoon cycles, each year some of the numbers above are multiplied by 2 or 3, depending on whether they are kept to first or second ratoon.

Computerisation of sugarcane breeding records was pioneered by the Hawaiian Sugar Planters' Association (HSPA) (Meyer *et al.*, 1974). The first BSES system was described by Hogarth and Skinner (1986, 1987), although CSR had a computerised system for pathology and breeding data called Canedata prior to this (Hutchinson and Roach, 1972). Most countries subsequently developed their own systems (e.g. Nuss *et al.*, 1989 for South Africa) and improvements were made to existing systems (e.g. Wu, 1987 for HSPA).

With the huge developments in computing power, software and web-based applications, there has been a gradual evolution from those initial systems. For BSES, this has involved a progression from an Image database system on an HP1000 mini-computer using Fortran programs for inputting data and for producing reports in the 1980s and some in-house data management software (DMAN) to a Microsoft Access Application (SPIDS—Sugarcane Plant Improvement Database System) and use of Microsoft Excel files from then up to 2003. However, in 2003, it was decided to develop a web-based, integrated plant improvement information system combining the previous less well-integrated systems used within BSES.

The initial step in developing the system was to involve all potential users to workshop what information was required and how it should appear. This involved breeders, pathologists and technicians as well as in-house and consultant database and software design experts. While this process has been ongoing since 2003, sufficient information was available initially to begin to design SPIDNet. The overall aims of SPIDNet were:

- 'everything in one place'
- integrated data capture, storage and retrieval
- standardised data collection operations
- validation of data using common rules

- secure and restricted data access
- automated routine operations
- increased accuracy and efficiency.

SPIDNet has been deployed as web, desktop and Windows Mobile applications. It uses Microsoft .NET Framework 2.0, C# programming language, Microsoft SQL Server 2005, ASP.NET for web development and Windows 2003 Server + IIS 7.0.

The main aim of this paper is to describe the major functionalities of SPIDNet and the more recent SPIDNet Mobile applications that have been developed to assist breeders, pathologists and technicians to improve the efficiency of their operations, reduce errors and facilitate data capture.

SPIDNet

The major functionalities of SPIDNet are briefly described below.

Manage plantings, plots and observations

A relational database typically comprises many 'tall and narrow' normalised tables. When these data are displayed or exported, they are usually pivoted or cross-tabbed into 'wide and flat' denormalised forms.

Lookup tables are used to record various 'Types' of data such as PlantingType or ObservationType—each with a unique TypeID. This approach is used for storing plantings, plots and plot observations in SPIDNet.

Once lookup tables have been populated, data can be loaded. First, plantings are defined and assigned a unique Planting ID. Plantings are aggregated into farms, mill areas and regions, each with its own lookup table.

Field plans are comprised of row, plot, plot type and genotype data, and are loaded into the Plot table, where each plot is assigned a unique Plot ID. Plots in the same field plan share the same Planting ID.

Observation values are then loaded into a PlotObs table referencing the appropriate Plot ID, Observation Type ID and crop class or harvest against which the observations were collected.

These data are then exported in special format for analysis by biometricians. The PlotObs table contains replicated, single plot data whereas analysed data is no longer specific to a single plot, so analysed data are stored in a separate table called PlantingObs. Raw plot observations and analysed planting observations share similar links to the same lookup tables. Using this system, a single observed value is linked to a specific GeneticEntity (GE), plot, planting and harvest.

SPIDNet now contains approximately 320 000 genotypes, 34 000 plantings, 500 000 plots, 4 000 000 replicated plot observations and 1 400 000 analysed observations. These cover breeding, selection, propagation and pathology programs, with plantings in the field, glasshouses and photoperiod facilities.

Pedigree management

The collective term 'GeneticEntity' or 'GE' was coined to refer to varieties, cultivars, experimental clones and crosses. Each GE is assigned a unique GeID and this GeID is referenced by Foreign Key relationships throughout the database, ensuring that all related data reference a known genotype.

Pedigree relationships are managed in the GeneticEntity table via a self-referential relationship between the ParentID and another row in the same table for the respective parent. This allows all pedigree relationships to be managed using only three columns in a single table.

Nomenclature variations are addressed through the use of synonyms. Each GE has a given name and may have one or more synonyms. When users upload data using alternative or outdated naming systems, SPIDNet transparently updates the old names to the newer names.

Each GE is summarised in a 'Passport' consisting of multiple tabs of related data (Figure 1). Wherever a GE name is displayed on a web page, it is hyperlinked to the GE's passport.

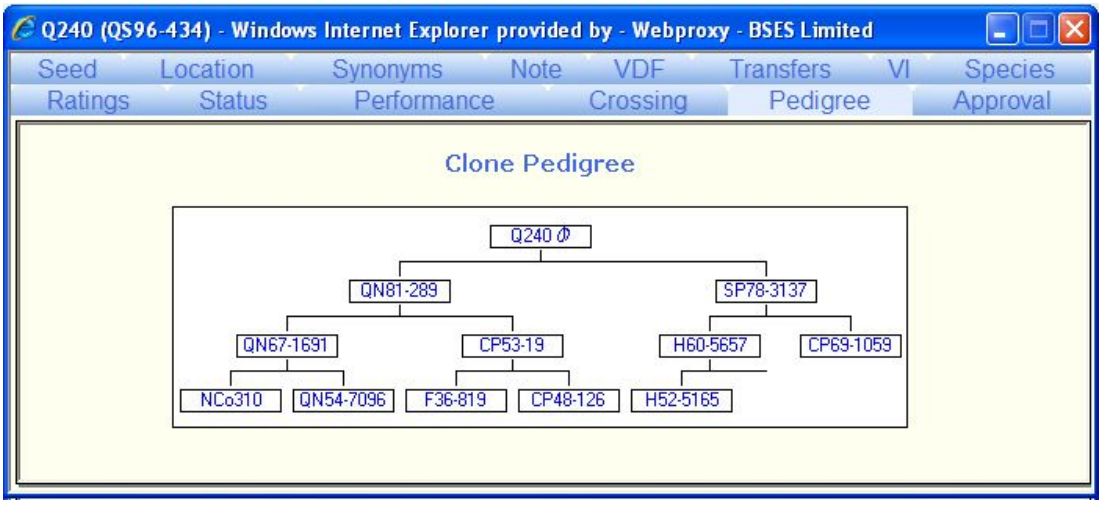

Fig. 1—SPIDNet passport for Q240$^{\Phi}$.

New GEs are added each year as either new crosses via the crossing program, or as clonal selections germinated from previous crosses. New clonal selections must be uploaded with their parents and the cross they were selected from. Parents of the cross and clone are cross checked and the import is rejected if anomalies are found.

Cross pollination

BSES maintains a collection of approximately 3000 active parents in both field and photoperiod facility plantings. Each year, a number of breeding objectives are established including Introgression, Foreign, Core and special purpose. Each

objective is assigned its own pool of parents to be crossed within constraints defined for the objective. Such constraints include inbreeding and violation of regional disease limits. Utilisation of available parents on a daily and seasonal basis is conducted according to targets established for each breeding objective.

A regional breeding value is derived for each parent based on individual and progeny performance and available disease data. When combined in a cross, these regional values are averaged to produce an overall cross status. The number of arrows to be used in each cross is governed by the quality of the cross and the number of regions the cross is suited to. The same regional cross status is used when selecting seed for the following year's Stage 1 trials in each region.

At regular intervals (usually 2–3 times per week) during the crossing season, a parent census is undertaken, where suitable flowers are located and counted, followed by an assessment of the pollen viability of flower subsamples. SPIDNet then examines each possible combination and processes them in conjunction with the constraints defined for the current breeding objective.

Proven crosses and parents are given priority based on selection success of past progeny. Until recently, a major deficiency of SPIDNet has been the lack of estimation of cross ratio, an important statistic formerly produced by the crossing register under the old crossing system. This has recently been rectified and will be used in the 2010 crossing season. Regional plant breeders have direct input by maintaining a list of priority crosses and parents to be adjusted as each census is published. Experimental parents new to the breeding collection are to be used in at least five crosses with established parents. Factorial polycrosses are used as the most efficient means of testing new parents. Biparental crosses are preferred for crosses using known parents; however, reciprocal crosses will be made between male parents having similar pollen strength.

Traceability of genotype movement

Misidentification of genotypes has been a more common problem in the past than breeders would like. If a particular genotype was sourced from an incorrect location, the mistake was often not uncovered until several propagation steps later. Similarly, if a disease outbreak was discovered in a propagation planting, it was often difficult to know the planting source and which other plantings came from this source.

To minimise the impact of these two scenarios, a traceability feature was added to SPIDNet. When field plans are loaded, each plot is assigned a unique plot ID. Traceability was achieved by requiring that each plot loaded must reference the source plot of the material being planted. Users must supply this source information as a Plot ID (often scanned from a bar-coded label), or as a unique location in the form of PlantingCode/Row-Plot (eg BIN05-03P/12-5). This self referential relationship means it is impossible to delete a plot if it has been used to source another plot. Traceability anomalies are detected and reported when users upload new field plans (Figure 2.)

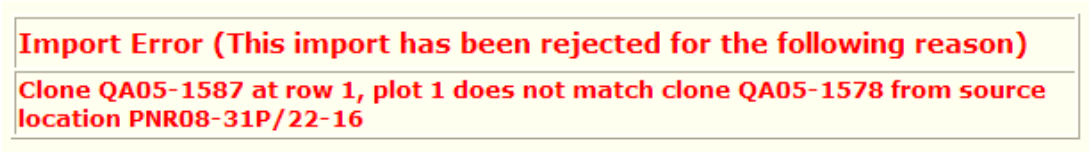

Fig. 2—Validation detects typographical error.

Sometimes source information is not available. To allow the traceability requirement to be circumvented, SPIDNet allows key users, such as the local plant breeder, to grant a number of plots to be loaded without source information.

Web services

Web Services are a means of exposing data and functionality to other software applications developed by BSES or their partners. Within BSES, web services are called by an Excel addin to add lookup capabilities from user's Excel files.

Other BSES web sites such as QCANESelect™ access SPIDNet functionality via web services to avoid duplicating features unnecessarily. For example, data displayed in regional variety guides and variety information sheets are entered into SPIDNet by Variety Officers. These data are then presented as a PDF report to SPIDNet users, and using web services, to the general public via QCANESelect™.

Other functions

A common requirement of any information system involves searching and retrieving data using key attributes. In SPIDNet, this means finding plantings and genotypes. Search results can be saved in custom lists and combined with other lists in advanced searches, or used as the basis for filtering data for display or export.

Reporting is another very important feature of SPIDNet which allows large amounts of data to be summarised in an easy to read format, usually in PDF format. The main reports are:

- Variety Data File (VDF), a comprehensive report including all selection trial results summarised over crops, trials, series and overall. Full disease data, available propagation material, current selection status, pending trial result summary, and pedigree are also included. These are used at annual regional Variety Selection Meetings where decisions are made on which clones to advance, hold or discard.
- Q Applications, a summary of all the relevant data on a potential new variety and is used to evaluate the case for release internally within BSES.

Storage and analysis of productivity data from mills is another feature of SPIDNet, providing information on regional varietal composition and allowing BSES to benchmark rates of genetic gain realised by the Variety Improvement program. Data from 1970 to the present are available.

SPIDNet Mobile

In 2007, a comparison of three popular barcode scanner equipped Portable Data Assistants (PDAs) was undertaken, and the Symbol MC70 was selected for BSES in-field data capture operations. These units employ the Windows Mobile 2005 operating system, which is fully compatible with Microsoft's .NET Framework and SQL Server 2005 Compact Edition products. Also supplied is a compact edition of Microsoft Excel—an application familiar to users and useful for data collection while custom applications were being developed. SQL Server's Merge replication feature was identified as a key to the seamless transfer of data between the main SPIDNet database and a smaller database located on the PDA.

Programming of the PDAs is relatively recent and the major applications developed to date are listed and described below. Further applications are planned.

Crossing operations

The first feature to be developed for the PDAs involved crossing operations. Parent plots were labelled with the unique Plot ID assigned when the planting was loaded into SPIDNet. On census days, users scan bar-coded plot IDs and record the number of suitable arrows available (Figure 3).

The flower sample is placed in a pollen book of bar-coded pages and the user scans the Page ID corresponding to the sample. In this way, the census count and flower sample are linked.

When all suitable parents have been visited, PDAs and pollen books are brought to the pollen testing laboratory. Here the samples are transferred to a bar-coded microscope slide and both Page ID and Slide IDs are scanned, effectively linking the census count to a microscope slide. Pollen viability is assessed for each slide, the microscope Slide ID is scanned and the pollen test value is scanned from a sheet of possible values. This links the pollen test result to the original census count and location.

Multiple PDAs are used during this process and, every 60 seconds, the data are sychronised to a central SQL Server database. This way, all data collected on any PDA are automatically copied to every other PDA—each PDA ends up with an identical set of data. When all tests have been conducted, the completed data are then used as the basis for crossing. Previously, all of the data were written by hand then manually entered into the SPIDNet database.

Harvest weights

Until recently, a field plan was printed as a grid of small cells into which a weigh truck operator would manually record each plot weight. This has now been automated using SPIDNetMobile, a Ranger 5000 weigh controller with RS232 interface, a Bluetooth to RS232 adapter, and a Zebra RW420 mobile printer. In 2009, plot weights were captured automatically for the first time in the BSES Variety Improvement program, improving efficiency and reducing the chance of manual transcription errors.

SPIDNetMobile contains all current plantings in a small database. When weighing a trial, operators select the appropriate planting, then choose one of several plot sort options. When the user has weighed a plot, they press the Print button on the weigh controller and the weight is transmitted to the PDA via the unit's RS232 port and the RS232 to Bluetooth converter. Upon receiving the data, SPIDNetMobile saves the weight against the current row and plot, then transmits the plot ID, row, plot, genotype and weight data via Bluetooth to the Zebra RW420 printer, thus maintaining a hard copy of the data as backup in the event of a PDA failure or loss. SPIDNetMobile then automatically advances to the next plot ready to receive the next weight (Figure 4).

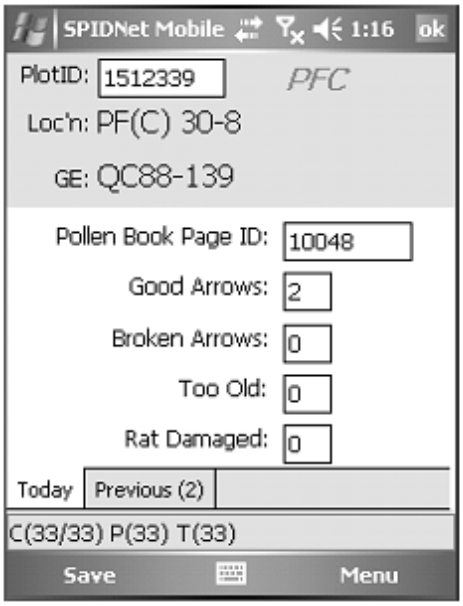

Fig. 3—Census data entry screen.

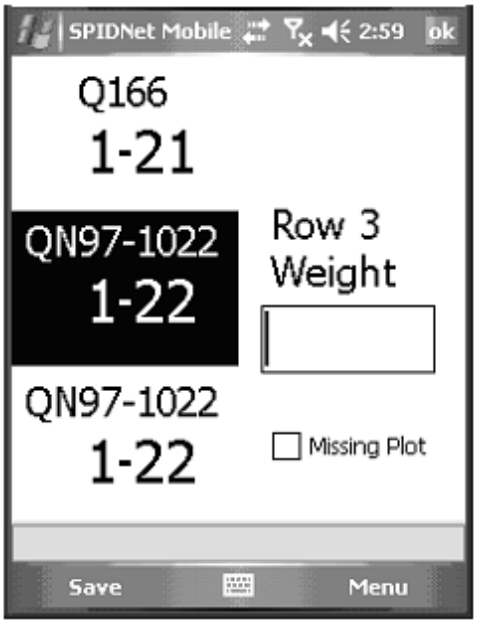

Fig. 4—Plot weight capture screen.

Capturing new plantings

The SPIDNet web site gives users access to a centralised label printing via a network of Zebra Z4M thermal transfer printers. When sourcing new material to be planted, users print a collection label for each genotype to be planted. These collection labels display the plot ID of the source plot as a bar-code.

Users define the new planting in SPIDNetMobile by supplying basic attributes including width and length (rows wide by plots long) and a planting code. Capturing new plot information for the new planting uses the same data entry screen as used for capturing plot weights, except that, instead of recording plot weights, it is used to record source plot IDs. When a plot ID is scanned, SPIDNetMobile is able to display the genotype corresponding to the source plot by looking it up in the corresponding

field plan. Some trials demand that a predefined experimental design is adhered to and users must plant particular genotypes in particular plots. To help enforce this requirement, SPIDNetMobile allows the use of a check file containing the expected clone for each plot to be planted. Users are notified of check file violations as they occur.

Data collection templates

Recently, it became apparent that a myriad of plot data could be collected if the data entry screens were available in SPIDNetMobile. The prospect of programmers producing and maintaining a multitude of data entry screens was obviously not a viable solution. To resolve this, the idea of Data Collection Templates (DCTs) was conceived. Users design DCTs via the SPIDNet web site, then collect data for a planting using the DCT in SPIDNetMobile. Text boxes, check boxes and list boxes are available to users when designing templates, each with an assignable default value. Text boxes are used for entry of alphanumeric data, check boxes are used for Boolean data, while List boxes can be populated with a list of items to be presented when collecting data.

To collect data using a DCT, users simply select a planting and a DCT (Figure 5), then SPIDNetMobile draws the required labels and controls in the order defined in the template, in the same data entry screen used for capturing plot weights and new plantings (Figure 6).

Fig. 5—Selecting a DCT.

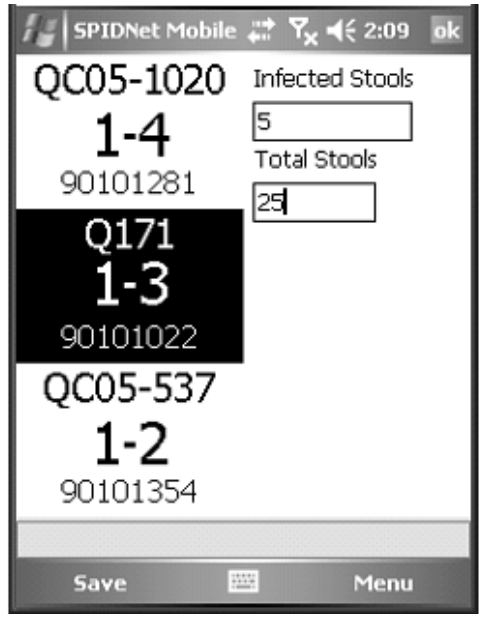

Fig. 6—Collecting plot data using a DCT.

A number of validation and shortcut features are available such as selecting list items by index using the numeric keypad, automatically advancing to the next

data entry item, automatically saving data then advancing to the next plot upon completion of the last data item, moving to the plot corresponding to a scanned plot ID and only enabling control for data entry if the scanned plot ID matches the plot ID of the current plot.

When finished, users can export the collected data from the internal database in .csv format. Columns included in all exports include plot ID, row, plot, clone, and a column for each observation in the template. Using DCTs, users have control over which observations they can collect with SPIDNetMobile without requiring any further software development.

Discussion

The integrated database, SPIDNet, has evolved significantly since development began in 2003. There has been a strong interaction between the field-based scientists and technicians, who are the users of the system, and the programmers, who not only are able to convert the ideas into computer language but also have in-depth knowledge of the science and operational basis of the breeding program. This has resulted in many refinements and additional functionalities that have added immense value to the variety improvement program. Such developments are certain to continue to further improve efficiencies and minimise errors. The recent development of SPIDNetMobile applications using PDAs has been an excellent example of this.

However, there are still many areas that need improvement and new functionalities that could be implemented to further improve SPIDNet. Some of these include:

- Auditing data to record who changed what and when.
- Storing raw pathology data (currently only analysed data are stored).
- Converting the 'SeedSelect' software from Access to a Desktop application in SPIDNet.
- Automating or semi-automating tasks such as selection of parents and crosses and managing the parent collection. Although systems have been developed to provide the data breeders require to make these decisions, it would be more efficient to program SPIDNet with the specific criteria breeders use to make these decisions. Work has commenced on managing the parent collection.
- Extending SPIDNet as a tool for all groups within BSES to manage trials (design and, treatments), upload raw data, export data for analysis and store analysed data.

Acknowledgements

We are grateful to key users for their ideas, patience, and willingness to test prototypes. These users include Alison Findlay, Richard Cervellin, Dion Appo, Ron Stevens, Heidi Clements, Vivien Dunne and Rhylee Pendrigh.

A significant amount of SPIDNet development has been contracted to external programmers. The authors would particularly like to thank Rebecca Watts for her tireless efforts and impressive capacity to translate seemingly chaotic systems into impressive software and data stores.

REFERENCES

Hogarth DM, Skinner JC (1986) Computerisation of cane breeding records. *Proceedings of the International Society of Sugar Cane Technologists* **19**, 478–491.

Hogarth DM, Skinner JC (1987) Computerisation of parental selection. In 'Copersucar international sugarcane breeding workshop'. pp 87–101. (Copersucar Technology Center, Piracicaba, Brasil).

Hutchinson PB, Roach BT (1972) Use of a computer in cane breeding. *Sugar Cane Breeders' Newsletter* **30,** 36–41.

Meyer HK, Heinz DJ, Lawrence N, Kimura E, Ladd SL (1974) Computer processing of sugarcane yield, breeding and selection records. *Proceedings of the International Society of Sugar Cane Technologists* **15**, 24–35.

Nuss KJ, Harding RL, Blose MJ (1989) The use of a computer in a sugarcane breeding programme. *Proceedings of the International Society of Sugar Cane Technologists* **20**, 963–967.

Wu KK (1987) Computer applications in sugarcane improvement. In 'Sugarcane improvement through breeding'. (Ed. DJ Heinz). (Elsevier, Amsterdam, Netherlands).