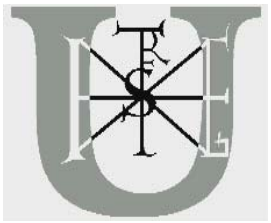

SZENT ISTVÁN UNIVERSITY

WHEAT BREEDING INFORMATION SYSTEM

MAIN POINTS OF THE PH.D. THESIS

CSABA KUTI

GÖDÖLLŐ

2007

Postgraduate School

Name: Postgraduate School for Plant Sciences

Field of Science: Crop Production and Horticultural Sciences

Head: **Prof. Ferenc Virányi, D.Sc.**
Department of Plant Protection
Faculty of Agricultural and Environmental Sciences
Szent István University

Supervisors: **Prof. Márton Jolánkai, D.Sc.**
Crop Production Institute
Szent István University

László Láng, D.Sc.
Head of Department
Agricultural Research Institute of the Hungarian Academy of Sciences

Approved by:

…….	……...	……
Dr. Ferenc Virányi	**Dr. Márton Jolánkai**	**Dr. László Láng**
Head of the Postgraduate School	Supervisor	Supervisor

Introduction

Over the last twenty years the use of computers has spread rapidly in agricultural research, and various techniques for handling data have evolved. International literature and experience in Hungary suggests that, for lack of special software, agricultural researchers, whether in small or large organisations, have been forced to find their own solutions. In most cases computerised data handling and evaluation has involved the use of word processors, spreadsheet programs, database handlers and statistical programs.

As the size of research projects has increased, usually with the optimisation of the human infrastructure, it has become obvious that this form of data processing is not suitable for the organisation of research tasks, the processing of large masses of data is difficult, it is impossible to check for correlations between non-structured data, decisions are uncertain when the necessary information is lacking, and problems are regularly encountered when preparing analyses and reports.

The need has arisen for a technology that integrates various types of data from complex sources, describes and assigns tasks, determines the order in which they should be carried out and

processes the results to provide data characterising the success of the research project.

In general it is still true today that while basic software can and should be purchased on the software market, user software and information systems satisfying the specialised requirements of agricultural research still need to be developed individually.

The thesis presents an information system designed to improve the efficiency of wheat breeding research in Martonvásár. The aim of the technology is to combine all the breeding and field experimentation data available to the research staff into a single system and to provide IT support for the various activities.

Aims

The need to combine all the research data and information into a single system backed up by IT support led to the expression of complex aims covering the whole range of activities involved in wheat breeding and field experimentation. The main objectives of developing the information system were as follows:

1. Design and execution of a uniform, integrated data structure
2. Transfer of input data to databases; development of automated data-collecting and manual input applications

3. Compilation of various lists and outputs with fixed and variable structure using data selected from the database
4. Organisation of complex field experimentation and breeding tasks
5. Incorporation of general data query functions
6. Development and supervision of comprehensive pedigree and gene bank records
7. Organisation of basic material exchange and registration of the relevant address list
8. Elaboration of a statistical module suitable for basic analysis, which could be run directly on data in the database
9. Provision of multiple concurrent user facilities

Materials and methods

Software

The Breeder application, developed in the Microsoft® Visual Basic Integrated Development Environment (VB-IDE), occupies approx. 86 MB, but does not need to be installed in individual work stations.

The majority of Breeder output and input data are in forms that can be displayed using the Microsoft® Excel, Microsoft® Access and Microsoft® Word applications of the Microsoft Office2000® or OfficeXP® packages, so the Office package should be installed at each work station to ensure optimum operation (approx. 150 MB).

Hardware

Work stations

These are the computers used by the developers and users (researchers, technicians) of the breeding information system and have a typical configuration consisting of Intel Pentium IV from 1.5 GHz (or compatible), memory: 256 MB RAM, hard disc: 160 GB, monitor: SVGA 17”, network card (10/100 Ethernet LAN), optical unit (CD/DVD), USB connection, mouse, keyboard.

Automatic data collectors

These are computers that are still operational, but whose performance level is insufficient for everyday use. Typical configuration: Intel Pentium III from 800 MHz (or compatible), memory: 128 MB RAM, hard disc: 40 GB, monitor: SVGA 15” (800×600), network card (10/100 Ethernet LAN), optical unit (CD/DVD), serial (RS232) connection, mouse, keyboard.

Network equipment

To ensure the servicing of requests from the work stations, a central database server of the IBM eSeries 200 type is operated with the following configuration: 1.6 GHz processor, 160 Gbyte hard disc, 1 Gbyte RAM.

Data files and storage requirements

The total size of the databases compiled each year from 1984 is approx. 200 MB. The files are not uniform in size, but are proportional to the increase in the breeding stock and in the type and number of data collected from year to year. In addition, simpler basic and parameter files of the text and Excel type, containing experimental designs, program settings and label forms, are also stored.

Applications based on barcodes

Label types

Barcodes are printed on commercially available, standard-sized self-adhesive or plastic labels. The self-adhesive labels are printed using conventional laser (HP, CANON) printers on standard Avery-Zweckform (AZ3651, AZ3652) label sheets, while the thermo-sensitive plastic labels are made of special material (HDPE) and are 0.25 mm thick, with a cross-shaped hole. They come in two sizes: 11 cm × 2.5 cm and 11 cm × 5 cm.

Barcode printers

The plastic labels are printed using special thermo-transfer printers (TTX300 COBRA, TTX350 OCELOT; Avery Dennison Corporation, 1985) with a maximum label width of 110 mm. From the technical point of view they differ considerably from conventional printers.

Barcode scanners

The measurement data are automatically read using optical scanners of the CCD (Charge Coupled Devices) type: BCH5X49, BS-L01.

Data-collecting devices

- NIR/NIT instruments: FOSS Infratec® 1241, Perten Inframatic® 8611
- Digital balances (Mettler) PB602-S, Viper SW, Spider SW
- Grain hardness meter: Perten SKCS 4100
- Dough expansion analysis (Chopin): Alveograph, Alveolink
- Measurement of rheological properties (Stable Micro Systems): TA.XTPlus
- Automatic data collectors fitted to combine harvesters: OMNIDATA Polycorder, HARVESTMASTER HM-Fieldbook

Statistical methods and experimental designs

Since replicated experiments carried out at a single location on the same field provide the most reliable estimation of the performance of a variety on the given field (Matuz, 1987), the statistical evaluation of replicated single-factor and multifactorial experiments was incorporated into the system. The statistical evaluation of a single-factor random block design and of various two-factor designs (random block, split-plot) is carried out and evaluated using the method described by Sváb (1981).

In the case of two-factor analysis of variance, the presence or absence of an interaction between the two qualitative traits may be important. If a significant interaction exists, the two-factor variance table is evaluated as suggested by Tóthné (Tóthné, 2004).

Among the breeding methods most frequently used in the breeding of self-pollinated cereal species (pedigree and mass selection methods; Bedő and Marton, 2004), the data structure was optimised for the pedigree method.

Results and discussion

An IT infrastructure was developed enabling the size and efficiency of breeding and field research programmes to be increased. A lucidly organised system was constructed, suitable not only for general use, but also reflecting the approach that has evolved in Martonvásár during long years of research.

1. Design and implementation of a relational data model

The Martonvásár wheat breeding data model was elaborated as part of the planned information system. The aim was to record all the data relevant to breeding in a uniform, integrated system, making it possible to classify and display the information and reports required for decision-making, to provide a link between breeding data and genealogical/gene bank data, and to keep a register of basic material exchanges and the relevant addresses.

Based on a flexible data model, comprehensive data integration was achieved, ensuring that all the data arising during breeding and field experimentation had a place in the database and could be adequately handled.

The data were grouped according to subject, leading to the following databases within the data model: breeding databases,

that are renewed from one year to the next; a genealogical database; a basic material exchange database; and a gene bank database. The number of tables found in each database is determined by the needs of the given subject. The breeding database, for instance, consists of 12 tables, containing data on experiments, field observation data, quality analyses, weight records, etc.

2. Organisational program modules

If the data are to be available within the information system, special applications are required to support various research tasks. These applications are based on the technological description of the task in question. The approximately fifty applications are included in the “Breeder” user interface, providing uniform availability to breeders and other agricultural staff. The uniform framework of the various modules contains menus, submenus and a quick lunch toolbar for the most frequently used program modules. The most important of these latter are applications assisting experimental design: the crossing, selecting and plot design modules.

When crosses are made the program automatically creates the new pedigrees (Purdy et al., 1968) and an information code

containing the selection history, and saves major data related to the various crossing programmes in crossing lists. These can later be used to check what the female and male components were, which experiments the parents can be taken from, and how many times a given parent has been used in the crossing programme.

The selection module records the origin of the genotype (previous year: experiment, plot) and the number of rows planted, automatically develops a code (selection history) and, if necessary, a new line identification number, and provides links with variety maintenance and frost and resistance testing. It automatically records the weight of the grain and, instead of individual designation and selection, it is possible to automatically select as many as several hundred genotypes according to previously adjusted scoring values and other conditions.

The plot design module is not only able to arrange the plots in various orders, but is also able to add a control plot in certain cases, while in other cases it may be an efficient tool for the automatic elaboration of experiments of a given size. This depends on which of the four plot design models available is chosen by the user. When the plot order is finalised the breeder has the option of distinguishing between lines with particular spike types, of arranging the lines in the order of heading in the previous year, of

grouping related lines within genotypes with the same heading date, etc.

The quick lunch bar can also be used to access the module designed to keep gene bank records, which permits data exchange between the gene bank and breeding databases (storage, data extraction). The module that stores information on all the incoming and outgoing seed lots of Hungarian and foreign breeding and gene bank basic material can also be accessed from here.

The frequently run modules include one that allows scoring data to be entered manually, a module that inserts the contents of external (Excel) files into the data structure of the information system, a module aimed at improving the consistency of genealogical data, a module that produces sowing plans and field books, a module for the design and printing of barcodes, and a module that automatically collects data from a wide range of laboratory quality analysis instruments and from automatic data collectors fitted to combines.

3. Pedigree and gene bank records

The cornerstone of breeding and field data records is a uniform pedigree model, capable of creating a unique identification code

for each genotype, and of handling the homonyms and synonyms that naturally arise during the development and use of genotypes.

A separate data structure was developed for these records, which currently contain the names of over 110,000 different genotypes. The pedigree data-handling program modules automatically assign two identification codes to each pedigree. The first (PAz) identifies the combination, while the second (SID) differentiates between sister lines arising from the same crossing combination.

The name of each genotype appears only once, in the central records, and genealogical information on the genotype can be extracted from here and attached to the records of individual years or experiments. For genotypes with several designations (pedigree, line name, variety name, etc.) the program modules handling the database decide which of these is required by the user in a given situation.

4. Data collection

When we speak of data collection we generally think of automated or semi-automated functions. The automation of data collection requires the application of the up-to-date techniques now widely available in the form of computer-linked, digitally

controlled measuring and analysing devices fitted with microprocessors. By automating data collection functions, the rate (and quality) of data collection and data processing can be brought to the same standard. This is extremely important in cases where time-saving is an important criterion (e.g. for the optimum exploitation of the short period between harvest and sowing).

Two things are of fundamental importance for automated data collection: communication with the relevant instrument, and automated data identification.

Communication requires an instrument-specific software interface, which must be created separately for each instrument or family of instruments. The work invested will be returned in full, as the rapid collection of accurate data will be ensured for the rest of the useful life of the instrument.

The basis for automatic data identification is the barcode, so a module designed to generate and print barcodes was incorporated into the system. This is capable of combining a number of data fields into a single barcode, and – very importantly – the data can be printed onto self-adhesive or plastic labels directly from the database.

The labels bear not only the barcodes, but also written information that can be used for scoring and seed preparation,

while the barcodes serve for the automatic identification of the data when they are entered in the databases.

5. Gene bank

The basic collection in the Martonvásár gene bank consists of 11–12 thousand lots. One important aim when planning the system was to distinguish gene bank data from breeding data, while allowing access in both directions. Lots from the breeding programme can be transferred to the gene bank for preservation, while gene bank lots can be exploited for breeding purposes.

In all cases the gene bank module makes selective suggestions on which genotypes to use for a given purpose. If new lots from a given experiment (or group of experiments) are to be chosen for gene bank storage, only genotypes not yet stored in the gene bank are displayed. If, on the other hand, we wish to know which of the genotypes from a given experiment could be used to freshen a given gene bank experiment, only genotypes already present in the gene bank will be displayed. The handling of data in the opposite direction, when gene bank lots are used in breeding, is extremely similar, with the important difference that gene bank management techniques are given more weight: lists can be prepared of gene bank lots that are more than a certain age or whose weight or

quantity (number of bags) is below a certain critical level. If a specific experiment is chosen, the module will indicate which critical gene bank lots could freshen using that experiment**.**

6. Statistical module

In some research programmes only a narrow time interval is available for the evaluation of the results. This is particularly true of breeding programmes, where reports must be prepared within days of harvest and only a few weeks are available for the preparation of the seeds that will be sown in the nursery in the next cycle. No special statistical packages are required for the basic statistical analysis of experiments sown at one or more locations with the same or different randomisation or for the preparation of the data for these analyses. The following basic statistical programs have been incorporated into the system:

- Descriptive statistics
  - mean indexes: numerical mean, median, geometrical mean, harmonic mean
  - variable data: maximum, minimum, sum
  - deviation indexes: variance, deviation, mean error deviation, coefficient of variance, mean confidence interval

- Randomised block analysis of variance
  - single factor
  - two-factor
- Split-plot design
- Linear regression analysis
  - simple
  - multiple
- Correlation matrix calculation

Main scientific results

1. The information infrastructure required to increase the size and efficiency of breeding and field research programmes was developed.
2. Comprehensive data integration was achieved, based on a flexible data model, thus ensuring the efficiency of the system for large-scale breeding and field research programmes. All the relevant data arising in the course of breeding and field experimentation have been assigned a place in the databases and all the tools required for handling these data are available.
3. An up-to-date pedigree recording system has been developed, allowing different combinations to be distinguished and lines of identical origin to be pinpointed. The designation of each

genotype is stored only once, in the central pedigree file, from which genealogical information on the genotype can be attached to the records of various years or experiments.

4. Program modules were developed on the basis of technological descriptions for the organisation of major breeding activities. These allow numerous work processes to be automated, and all the previously recorded information on each genotype to be accessed when making decisions on selection, thus creating a network for data handling and team work.
5. The conditions were created for automated data collection. With the help of a built-in barcode generator, codes were developed to facilitate automatic data identification, and specific data collection programs were written, capable of communicating with the recording instruments.

Conclusions and recommendations

If wheat breeding research is to remain competitive, it is important to be able to expand the size of the research programme, to handle the vast array of data and, above all, to synthesise the information essential for decision-making from the millions of data generated by a major research programme from year to year.

The thesis describes the elaboration of an information system for this purpose. In addition to a presentation of the results, a survey is also given of the software currently available for use in crop production and plant breeding.

When elaborating the information system a basic assumption was that teamwork was required for a large-scale research programme, with several people working simultaneously but intermittently on each project. This meant that considerable freedom had to be built into the system, to allow the scientists and technical staff to carry out joint activities at different times. The system must be able to keep track of all the changes and allow tasks to be continued where they were left off, avoiding the need to start anything again from scratch. This explains one of the major differences between this system and the other software examined. In the latter, a number of parameters have to be given right at the beginning, which are still not available when

genotypes are being selected for a large-scale research programme, or which could only be given at this stage with a considerable effort.

The flexibility of the data structure and the way in which the individual tasks are organised makes it possible to start certain tasks before other parts of the programme have been finalised. For instance, it is possible to start generating barcodes for selected genotypes and using them for seed preparation before all the genotypes required for a given experiment have been selected from the thousands of genotypes involved in a previous experiment.

The handling of genealogical data differs greatly in the various types of software examined, depending on the quantity and complexity of the available genealogical data. The greater the quantity and variability (different species) of the genealogical data processed by a given system, the more complex and important handling them adequately becomes. The construction and handling of genealogical data plays a central role in the *Breeder* program, in terms of both the number of data available and the complexity of the handling methods. In contrast to all the other types of software investigated, the basic material records required for the execution of large-scale breeding and field programmes are closely linked to the data stored in the gene bank.

The aim of the statistical module of the *Breeder* program is to ensure the rapid evaluation of breeding and field experiments by allowing statistical programs to be easily run directly on the measuring and observation data stored in the databases. No special statistical packages are required for the basic statistical analysis of experiments sown at one or more locations with the same or different randomisation or for the preparation of the data for these analyses.

Unique items in this system are the incorporation of the barcode technique and the widespread use of automated data collection using these codes; the data query functions, consisting of a large number of tables covering many years, which ensure the informative nature of the system; and the printing of sowing lists and field books with fixed or variable form and content.

Through the introduction of the computerised data processing and task organisation outlined in the thesis, cost-, time- and labour-intensive work processes have been simplified, while minimising errors resulting from inaccuracies and inconsistencies. In this way the system has become an important tool for improving efficiency and competitiveness.

References

Bedő Z., Marton Cs. (2004): Plant breeding methods. pp. 71–100. In: Bedő Z. (ed.): *The Birth of the Seed. Theory and Practice of Seed Production*. Agroinform Kiadó, Budapest, 537 pp.

Matuz J. (1987): Importance and problems of farm-scale experimentation. pp. 223–236. In: Barabás Z. (ed.): *Manual of Wheat Production*. Mezőgazdasági Kiadó, Budapest, 223-236.

Purdy H. L., Loegering W. Q., Konczak C. F., Peterson C. J. Allan R. E., (1968): A proposed standard method for illustrating pedigrees of small grain varieties. Crop Science, 8: 405-406.

Sváb J. (1981): *Biometrical Methods in Research*. Mezőgazdasági Kiadó, Budapest. 557 pp.

Tóthné L. K. (2004): Statistical hypothesis analysis. pp. 30–47. In: Szűcs I. (ed.): *Applied Statistics*. Agroinform Kiadó, Budapest, 30-47.

Scientific publications serving as the basis of the thesis

Papers published in reviewed journals

Láng L., **Kuti Cs.**, Bedő Z. (2001): Computerised data management system for cereal breeding. Euphytica, 119.1-2: 235-240.

Marton L.Cs., **Kuti Cs.** (2002): Modified joint scaling test for evaluating the effect of level of heterozygosity of the female parent. Növénytermelés, 51.4: 1-6.

Marton L.Cs., **Kuti Cs.** (2002): Modified joint scaling test for evaluating the effect of level of heterozygosity of the female parent. Acta Agronomica Hungarica, 50.2: 185-190.

Kuti Cs., Láng L., Bedő Z. (2003): Computerized recording of mass measurment data from field experiments. Növénytermelés, 52.3-4: 329-340.

Kuti Cs., Láng L., Bedő Z. (2004): Use of barcodes and digital balances for the identification and measurement of field trial data. Acta Agronomica Hungarica, 52.4: 409-419

Bedő Z., Láng L., Veisz O., Vida Gy., Karsai I., Mészáros K., Rakszegi M., Szűcs P., Puskás K., **Kuti Cs.**, Megyeri M., Bencze Sz., Cséplő M., Láng D., Bányai J. (2005): Breeding of winter wheat (Triticum aestivum L.) for different adaptation types in multifunctional agricultural production. Turk. J. Agric. For., 29: 151-156.

Kuti Cs., Láng L., Bedő Z.(2006): Pedigree records in plant breeding: from independent data to interdependent data structures. Cereal Research Communications, 34.2-3: 911-918.

Klara Meszaros, Ildiko Karsai, **Csaba Kuti**, Judit Banyai, Laszlo Lang, Zoltan Bedo (2007): Efficiency of different marker systems for genotype fingerprinting and for genetic diversity studies in barley (Hordeum vulgare L.). South African Journal of Botany 73: 43-48.

Conference proceedings, conference abstracts

Láng L., **Kuti Cs.**, Bedő Z. (2000): Computerised data managerment system for cereal breeding. 6th International Wheat Conference 5-9 June 2000, Budapest, Hungary, Abstracts, 91

Láng L., **Kuti Cs.**, Bedő Z. (2001): Computerised data management system for cereal breeding. Wheat in a Global Environment Proc. of the 6th International Wheat Conference, 5-9 June 2000, Budapest, Hungary, Kluwer Academic Publishers 561-569.

Láng L., **Kuti Cs.**, Bedő Z. (2002): Martonvásár software in cereal breeding. Chinese-Hungarian Workshop on "Molecular Genetics and Breeding in wheat". Martonvásár, Szeged 21-26 May, 2002 Lecture Abstracts, 11-12.

Láng L., **Kuti Cs.**, Bedő Z. (2003): Computerised recording and analysis of plant breeding data using the “BREEDER” program package developed in Martonvásár. IX. Növénynemesítési Tudományos Napok, Összefoglalók, 2003. Márc. 5-6. MTA Budapest, 18.

Kuti Cs., Láng L., Bedő Z. (2004): Identification and measurements using barcodes and digital balances.. X. Növénynemesítési Tudományos Napok, Poszter Összefoglalók, 2004. Február 18-19, MTA Budapest, 125.

Kuti Cs., Láng L., Bedő Z. (2005): Pedigree records in plant breeding: from unordered data to a uniform data structure. XI. Növénynemesítési Tudományos Napok Összefoglalók, 2005. márc. 3-4. MTA, Budapest, 47.

Bányai J., Szűcs P., Karsai I., Mészáros K., **Kuti Cs**., Láng L., Bedő Z. (2005): Variety identification at the DNA level in wheat.. XI. Növénynemesítési Tudományos Napok Összefoglalók, 2005. márc. 3-4. MTA, Budapest, 60.

Other publications

Bedő Z., Szunics L., Láng L., Veisz O., Karsai I., Juhász A., Rakszegi M., Vida Gy., Szűcs P., **Kuti Cs.**, Megyeri M., Gál M. (2002): Breeding. Annual Wheat Newsletter, Vol. 48:66-70.

Bedő Z., Láng L., Szunics L., Veisz O., Vida Gy., Karsai I., Mészáros K., Juhász A., Rakszegi M., Szűcs P., Puskás K., **Kuti Cs.**, Megyeri M., Gál M., Nagy IJ. (2003): Items from Hungary, Department of Wheat Breeding, Agricultural Research Institute, Martonvásár. Annual Wheat Newsletter, 49:30-34.