Pedigree records in plant breeding: from independent data to interdependent data structures

C. KUTI – L. LÁNG – Z. BEDŐ
Agricultural Research Institute of the Hungarian Academy of Sciences, H-2462 Martonvásár, POB 19, Hungary
corresponding author email: kutics@mail.mgki.hu

Summary

The storage of wheat data in computers began in the mid-eighties in Martonvásár, and was accompanied by the development of the first simple programs to assist the data management of routine breeding tasks. The great expansion of breeding materials and the demand for new applications have led to an enormous increase in the number of data and have made data processing increasingly more complicated. Data storage facilities and computer programs reflecting an outdated technological level were unable to keep pace with developments. Data storage and applications had to be redesigned on new lines to create a completely new information system amalgamating know-how from breeding and informatics.

The paper introduces an extremely important part of this system: pedigree records, which contain the designations of all the genotypes included in traditional field breeding programmes and in the gene bank, together with crossing data, phenotypes and genomic data.

An up-to-date, consistent pedigree register is one of the key components in the breeding information system, without which the maintenance and alteration of the names of plant species (wheat, barley, oats, etc.) and linking them to experiments and experimental quality data would be an extremely complex, time-consuming task. It would be even more difficult to keep track of all the genotypes and the increasingly large numbers of related lines from year to year.

In addition to describing the rationale behind the system, details will be given on the tools and conditions required for the establishment of the pedigree records, and the internal and external sources available. Finally, some practical examples will be given of how the Martonvásár wheat breeding information system has been applied.

Key words: agroinformatics, pedigree, breeding, software, applications

Introduction

One of the main tasks facing the Martonvásár research institute is to develop high quality wheat varieties. The breeding process, from crossing to the registration of new varieties, generally takes 8–9 years and involves several consecutive cycles. Each cycle of a large-scale breeding programme, which coordinates the work of numerous researchers and technical staff, generates an immense volume of data from field observations, complex analyses and feed-back evaluations. In general these data are generated and stored independently, i.e. all the participants attempt to handle the data arising in the course of their work within their own informatic infrastructure (own computer, software, network).

Eventually this inevitably leads to the same data being stored under different names, making the extraction and handling of the information slow, the data inaccurate and the establishment of links between the data sets almost impossible.

The primary aim in setting up a local information system is to make the information stored on the hardware (computers) available to the user rapidly and accurately with the aid of software and other applications.

DOI: 10.1556/CRC.34.2006.2-3.219

Over the last ten years the use of computers has spread rapidly among people working on agricultural research. During this period over 40 papers on breeding-linked software developments were published in the Agronomy Journal (American Society of Agronomy), 69 in Crop Science (American Society of Agronomy) and 124 in Plant Breeding (Blackwell Publishing). Most of these papers were concerned with plant modelling or gene technology. In the four journals reviewed, only one (Euphytica) contained a mention of a data model and relevant applications for conventional breeding (*Láng et al.*, 2001).

Naturally, not all information on computer developments for breeders is to be found in scientific journals. Examples of this are the software developed by Agronomix Software Inc. (Agronomix, 2001) or the information system developed jointly by CGIAR, NARS and ARI (ICIS, 2000). The TrialWizard software developed in Holland (Beerpoot Consultancy BV, 2004) is considerably less complicated than these and can be used to design and evaluate very simple experiments. Agrobase may attract interest due to the wide choice of experimental designs and statistical analytical tools it offers. The Pedigree Data Management Module, useful in designing experiments, consists of three main parts: the Parents module, a module containing tools for designing field trials and crosses, and the Population module.

The Parents module contains the names and characteristics of the male and female parents available to the user. Each is given a name for use within the system (not identical with the pedigree) and these can be selected for field experiments or for use in new crosses. When a male or female parent is selected from the Parents module, details are provided of the crossing history, including the date of crossing and the researcher who made the cross.

Experiments can be designed using the design module. All the information required for the establishment of the experiment can be given in seven steps. These include the name and type of the experiment, the experimental design and size, the genotypes and experimental locations to be included, and the traits to be tested. Special features can also be programmed. This module can also be used to create new populations (Fn), which must be given code names. If crosses are to be made it is possible to select the breeding system (self-/cross-pollinated) and the form of the pedigree (American, European).

The third component in the Pedigree Data Management Module is the Populations module. This acts as an Fn warehouse and transition station for lines being transferred between generations, where the populations can be grouped and converted. The main purpose of converting a population to a parent is when the population is no longer an experimental line and has become a registered variety, involving the danger of losing the associated selection history.

The advantages and disadvantages of this system of pedigree records can be summarised as follows: various different experimental designs can be chosen, almost anything can be grouped (experiments, parents, populations), crosses can be traced using the crossing history, the traits to be tested in the experiment can be chosen for each individual case, and sowing lists and observation records can be printed out. It should be noted, however, that the system only works efficiently for the registration of experiments and data from a moderately sized breeding programme (up to a few thousand genotypes). As there is no annual breakdown of the experiments, it becomes difficult to plan experiments when selection is made from year to year in large-scale breeding programmes. It is time-consuming to go through the whole stock of genotypes, or a large part of it, when selecting parents and populations. Very little is done automatically, so a lot of manual work is involved. Some of the functions are slow even for a low number of (demo) data.

The main aim of the ICIS database system is to facilitate the uniform use of genealogical and pedigree information for many plant species, arising from a number of sources. The data model can be established for any plant species using a standardised design. This consists of two separate databases, a central database and a local database. Every possible trait related to the genesis, genealogy, nomenclature and chronology of the genotypes of major plant species are recorded with the help of 16 tables in the central database. Within each plant species, these are uniform for all installations and can only be modified by the central ICIS administrator. The system can be individualised by each user with the help of the local database, the structure of which is only slightly different from that of the central database, with the important difference that it can be modified by local users. The system is such that the users do not perceive that they are using two separate databases. The 16 tables in the central database contain the following groups of data: four tables contain the pedigree data, three look-up tables supervise references, method codes and the fields defined by the user, three others supervise the users and installations, while the remaining six contain detailed data on the sites of origin of the genotypes.

This provides a very thorough, detailed modelling and registration of genealogical data. Containing as it does international data on a large number of plant species, it is too complex and far-reaching for those dealing with a single plant species, though naturally the genealogical data it contains are of great value for all the breeders of wheat, maize, rice, etc. to whom it is accessible.

In comparison with the systems outlined above, the TrialWizard is a much more modest system, capable of storing all the data required for a single experiment, from the names of the genotypes to the observation data, in a single file. There is no question here of either a data model or a separate pedigree record.

The design and supervision of the consecutive experiments required each year for the development of a new wheat variety can be ideally achieved using the information system developed in Martonvásár (*Láng et al.,* 2001) thanks to the specific applications and the relevant pedigree records.

Materials and Methods

Data model

Data modelling (Szelezsán, 1998; Date, 1982) is an up-to-date method for the design of data files. If a company or institution establishes and operates a database handling system, all the available data will become a central resource. If the data are carefully entered into a uniform system, each piece of information will form a single entry, and all the entries will be in uniform, compatible form. The database only becomes of value to the user if it is equipped with software (programming language) that allows it to be handled. The extraction of data from the databases is achieved using software or applications at several levels. The top level is the application actually available to the user for the creation, deletion or modification of the data. A general application development program designed by Microsoft® was used to create the application. The use of Visual Basic® (*Jamsa and Klander* 1998, *Aitken* 1999) ensured the Windows (Microsoft, Windows 2000) surface familiar to the majority of users. This top level application uses the syntax of Structured Query Language (SQL – *IBM®,* 1974) to compile the structured queries that can be interpreted by the next level of software, which handles the database and is capable of using the operating system to extract data from the hard disc and load them into the computer.

The databases available to the user come into existence as the result of a multistep process. Without listing or detailing all these steps, only the most important will be discussed here. The first step is to take stock of requirements and understand what is involved. This is followed by the establishment of data tables, creation of links between these tables, the normalisation of the data, the organisation of the files, indexing, screen design, etc.

Software specifications and hardware requirements

In choosing the operation system and the software development tools, an important consideration was the rational utilisation of the existing computer background (hardware and software). Care was taken to ensure that the applications could be run on all computers using any Windows® (Norton et al., 2000) operation system (WIN9x, NT, 2000, XP). The amount of hard disc storage space required (approx. 300–400 MB) naturally depends on the given volume of data, but this is unlikely to be a problem, as most computers now have a storage capacity of 60–80 GB.

Network

For computers that are part of a network infrastructure, extremely high performance can be achieved. There is a hardly perceptible reduction in the speed of data handling, which is completely offset by the fact that the databases can be used simultaneously by several users. In this case the data files do not take up space on the PCs, but are stored on the central server and data input is immediately available to all users on the network.

Pedigree nomenclature

After establishing the physical skeleton of the pedigree registration system, care must be taken when loading and using data to adhere to the rules governing the uniformity of pedigree nomenclature and the distinctness of the designations used.

The method proposed by Purdy et al. (1969) was taken as the basic principle. This is an adaptation of the method elaborated by Wiebe (1961), replacing the symbol X by / (primary cross), 2X by // (secondary cross) and 3X by /3/ (tertiary cross), etc.

The nomenclature for backcrosses has also changed. Instead of using a superscript next to the cross symbol (X) on the side of the recurrent parent, e.g. Sky X^2 Lark, the number of backcrosses is written on the same line as the rest of the pedigree, with an asterisk to denote the recurrent parent, e.g. Sky / 2*Lark. Some of the symbols proposed by Purdy were not used in the present system, for example the commas recommended to distinguish between related lines selected from a cross (e.g. "THATCHER, MARQUIS / KANRED // IUMILLO / MARQUIS,CI10003,MINN 2303,NSN II-21-23"), since these could be confused with the commas used in the syntax of the query language (e.g. SQL).

In aiming at distinctness in pedigree nomenclature it is important to eliminate anomalies caused by genotypes (combinations, varieties, lines) being listed in the data records under several names. In order to solve this problem, international technical databases were used in addition to local sources. The most important of these was the wheat pedigree information contained in the Central Database of the Genealogy Management System (GMS) set up by CIMMYT (ICIS, 2000).

The online "Web" database (WPIAG) established jointly by the Russian Academy of Sciences and the Czech Crop Production Institute, which contains over 69,000 pedigrees from 2,529 sources, together with information on numerous identified alleles, also proved extremely useful.

Another important reference was the over 14,000 wheat pedigrees published by Zeven et al. (1976).

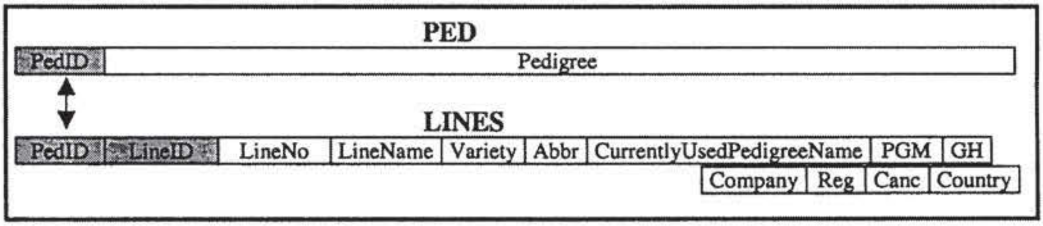
Results and Discussion

Data structure

In the first step all the pedigree data available in various files were collected in a single large file containing several tens of thousands of items. These pedigrees were thoroughly checked to eliminate misprints and errors, after which a special chaining program was used to identify related lines (sister lines). All distinct combinations were given a code number (integer), designated as the pedigree code (PedID). Related lines belonging to each combination were supplied with a subcode within the PedID code, designated as the line code (LineID).

The "Ped" and "Lines" tables located within the PedLines database, completely separate from the experimental and observational data, are major components in the pedigree registration system (Table 1).

Table 1. Column structure of the Ped and Lines tables



PED

PedID	Pedigree

LINES

PedID	LineID	LineNo	LineName	Variety	Abbr	CurrentlyUsedPedigreeName	PGM	GH	Company	Reg	Canc	Country

Columns in the Ped table:

PedID: a single code is generated for each pedigree, e.g. 36041

Pedigree: the pedigree itself (where known), e.g. MV21-85/MV15

Columns in the Lines table:

PedID: code of the combination in the Lines table

LineID: to distinguish between related (sister) lines within the combination

LineNo: in-house used distinctive number (in the early stages)

LineName: more advanced designation used to distinguish between lines

Variety: names of state-registered varieties

Abbr: abbreviation of the variety name, used chiefly in crosses

CurrentlyUsedPedigreeName: the most advanced designation currently used for the genotype

PGM: plant species

GH: growth habit

Country: abbreviation of the country where state registration took place

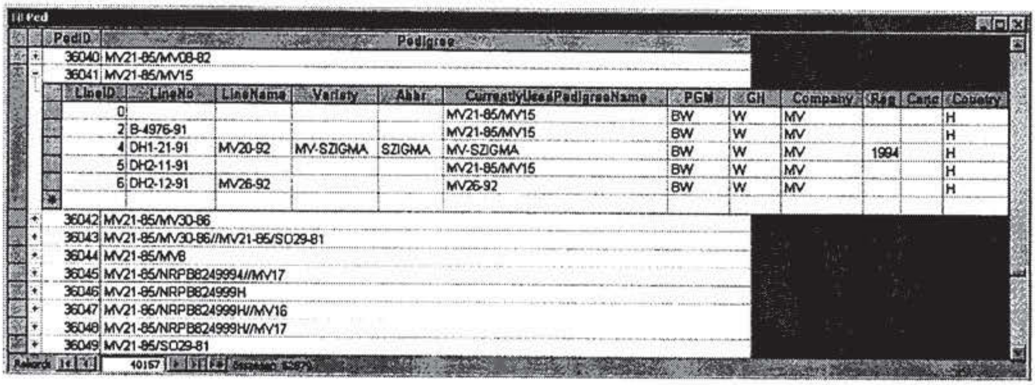
Company: name of the company entering the genotype for state registration

Reg: year of registration

Canc: year when registration was cancelled

Grey fields are key fields whose contents can only be found once in the whole table. The type of relationship between the two tables (One-to-many) denotes that one pedigree from the Ped table may be linked with many related lines from the Lines table, as illustrated in Figure 1.

Figure 1. Related lines within a combination



PedID	Pedigree
36040	MV21-85/MV08-82
36041	MV21-85/MV15
36042	MV21-85/MV30-86
36043	MV21-85/MV30-86//MV21-85/SO29-81
36044	MV21-85/MV8
36045	MV21-85/NRPB8249994//MV17
36046	MV21-85/NRPB824999H
36047	MV21-86/NRPB824999H//MV16
36048	MV21-85/NRPB824999H//MV17
36049	MV21-85/SO29-81

LineID	LineNo	LineName	Variety	Abbr	CurrentlyUsedPedigreeName	PGM	GH	Company	Reg	Cert	Country
0					MV21-85/MV15	BW	W	MV			H
2	B-4976-91				MV21-85/MV15	BW	W	MV			H
4	DH1-21-91	MV20-92	MV-SZIGMA	SZIGMA	MV-SZIGMA	BW	W	MV	1994		H
5	DH2-11-91				MV21-85/MV15	BW	W	MV			H
6	DH2-12-91	MV26-92			MV26-92	BW	W	MV			H

Naturally, the pedigree registration system contains more than these two tables. Important data are also available in other data structures, such as crosses in the "Cross" table, phenotypes in the "Pheno" table, genes and markers in the "Genes" table, parental ancestors for the preparation of Mendelgrams in the "GenAnal" database, etc.

When planning the structure of the "Ped" and "Lines" tables, which are independent of each other but closely linked, the following aspects were considered:

- individual identification on the basis of "PedID" and "LineID"
- easy grouping and distinguishing of related lines
- single registration of each piece of pedigree information within the system
- network operation to allow multiple simultaneous users
- data input and data extraction in familiar forms, e.g. Excel

Data traffic within the pedigree register

The pedigree register is handled by the modules of the Breeder application package. It is also possible to establish pedigree records directly, but the large majority of new entries are registered indirectly in the course of breeding activities. This is particularly true of the automatic registration of combination names when new crosses are made, of the registration of information on new seed lots, and of the creation of new codes and names to distinguish new lines established during selection.

Basically, the pedigree register is a continually expanding database, but tools are available to identify superfluous records and to propose their deletion.

Incorporated functions

With the help of the incorporated functions to be found in the menu, very useful information can be obtained within an extremely short time.

After activating the required menu point, if the user begins to type in the name of the desired combination, variety or line, the system automatically suggests names from the register whose first letters correspond to those typed in. Then, instead of typing the whole name, it is possible to scroll down to the desired name on the list. After clicking on this name, the whole designation becomes visible (for example, if a variety name is entered, the combination, line and abbreviated name are presented). By clicking on the search button, full information can then be obtained on what crosses and experiments the given genotype was used in over the last 20 years and on where it is listed in the gene bank register.

After a similar process of selection, a search can be made for the ascendants of the desired genotype, and a list will be obtained of all the varieties or genotypes involved in the pedigree, including percentage shares. Naturally, the more data on pedigrees are available in the system, the more informative this list will be, so it is important to constantly expand the pedigree register using descriptions from other databases.

Specific data

Wheat genotypes have certain traits that do not change in consecutive stages of the breeding process, such as spike type. This is not true, however, of the designations created using the Purdy nomenclature when crosses are made. In later stages the complicated combination names may be replaced by simpler names giving a better reflection of the breeding stage. Naturally, breeders would like the new names to be valid in retrospect, too, without deleting the previous names. This has been solved by creating a special data field in the Lines table in the applications containing pedigree names, which always contains the most up-to-date name of the genotype, and care has been taken that at any stage in the breeding process, if the name is changed, the contents of this field will automatically be updated.

Other possibilities

- As pedigree names only appear once in the data model, it is far easier to carry out maintenance on the register than it was when each name could have been included in thousands of files. One of the most important maintenance functions is consistency checking. This involves commanding the system to go through the annual databases and check that all the genotypes used in all the experiments have a valid, linked pedigree name. This process can also be reversed to check whether there are any entries in the pedigree register that do not correspond to any genotype in any experiment ("empty" names), which the system then suggests deleting.
- It follows from the structure of the data model that related lines belonging to the same combination can easily be grouped. It is very useful in the course of selection to be able to review where a genotype chosen from any particular experiment has been used in the whole of the breeding programme in the given year, as this may enable the breeder to include the genotype in next year's material from a different experiment.
- The connection between the annual breeding programme and the gene bank also comes about through the pedigree register. Only genotypes from the annual breeding stock that are not yet to be found in the gene bank are suggested for inclusion. On the other hand, when gene bank accessions are chosen for refreshing, the genotypes in the breeding material that are already present in the register are listed.
- Another useful property of the system is that Excel files can be used for all grouped input and output. For example, such files can be used to enter pedigrees or to print out Mendelgrams, crossing lists, etc.
- The use of a pedigree register containing unmistakable names makes it possible to publish information on certain genotypes, together with their most characteristic quality data, on the institute's web page.

In addition to the applications listed here, all the program modules that use genotype names are also linked to the pedigree register.

The use of computers in institutes dealing with agricultural research has grown substantially over the last fifteen years. This has naturally led to endeavours to create software, applications and information systems designed to assist research activities. The new, integrated pedigree register set up in Martonvásár contains separate tables including over 30,000 combinations and the almost 100,000 lines related to them.

The advantages of this pedigree register, which can be accessed using the applications included in the system, have been enjoyed by the scientists and other staff of the breeding department for over four years. The system was developed by wheat breeders, information scientists and other technical staff. The menus and toolbar functions of the program are in English language.

References

Agronomix Software, Inc. (2001): Agrobase Generation II, 171 Waterloo St., Winnipeg, Canada.

Aitken, P.G., (1998): Visual Basic 6 Programming Blue Book. The Coriolis Group.

American Society of Agronomy: Archive of All Online Issues: 1 Jan 1998 - 28 Mar 2005. Crop Science, CSSA Headquarters Office, 677 S. Segoe Road, Madison, WI 53711.

Beerpoot Consultancy BV (2004): TrialWizard, Hartenseweg 32, 6705 BK Wageningen, Holland.

Blackwell Publishing Oxford UK, Blackwell Verlag GmbH, Germany: Plant Breeding

Bott, E., Leonhard, W. (1999): Special Edition Using Microsoft Office 2000. Que Corporation.

Date, C. J. (1982): An introduction to database systems. Addison-Wesley Publ. Co., Reading, MA.

International Crop Information System (ICIS), (2000): Technical Development Manual, CIMMYT (Centro International de Mejoramiento de Maiz y Trigo) Mexico, IRRI (International Rice Research Institute) Philippines.

IBM, (1911): International Business Machines Corporation, New Orchard Road Armonk, NY 10504.

Jamsa, K., Klander, L. (1997): 1001 Visual Basic Programmer's tips. Jamsa Press, Las Vegas. USA.

Láng L., Kuti C., Bedö Z. (2001): Computerised data management system for cereal breeding Euphytica, vol. 119, no. 1-2, pp. 235-240(6) .

Microsoft Corporation: Access 2000. Microsoft Corp., Redmond, WA.

Microsoft Corporation: Microsoft Windows 2000 Professional. Microsoft Corp., Redmond, WA.

Norton, P., Mueller, J., Mansfield, R. (2000): Complete Guide to Microsoft Windows 2000 Professional, SAMs Publishing, Macmillan USA.

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, vol 8, july-august.

Szelezsán, J. (1998): Adatbázisok. LSI Oktatóközpont, Budapest.

Wiebe, G. A. 1961: On writing complex hybrids in small grain breeding. 1960 Barley Newsletter 4:13-14.

VIR St. Petersburg N.I.Vavilov Institute of General Genetics RAS, RICP Praha: Wheat Pedigree and Identified Alleles of Genes (WPIAG).

Zeven A. and Zeven-Hissink NC (1976): Genealogies of 14000 wheat varieties. The Netherlands Cereal Centre, Wageningen, International Maize and Wheat Improvement Center, Mexico. 121 pp.

Received 26 August, 2005, accepted 21 November, 2005