



UPOV/DATA/BEI/04/6 Rev.

ORIGINAL: English

DATE: June 2, 2004

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

WORKSHOP ON DATA HANDLING

organized by
the International Union for the Protection of
New Varieties of Plants (UPOV)

in cooperation with
the State Forestry Administration of China,
the Ministry of Agriculture of China and
the State Intellectual Property Office of China

with the financial assistance of
the Ministry of Agriculture, Forestry and Fisheries of Japan

Beijing, June 9 to 11, 2004

DATA CAPTURE AND STORAGE OF DATA

Document prepared by an expert from Germany

DATA CAPTURE AND STORAGE OF DATA

Uwe Meyer, Bundessortenamt, Hanover

I. Introduction

1. Data from measurements and visually assessed characteristics can be recorded electronically using a handheld computer (“datalogger”) or on a paper form. They are then transferred or entered into an electronic form of tables (files or database tables) in a personal computer (PC). Following this, the data must be checked and analysed after which the crop expert can produce the variety descriptions. From the data capture in the field to the production of the variety description, the crop experts are responsible for the correctness and security of the data.

II. Trial plan and structure of data

2. As a first step, the crop expert should establish the plan of the field trial with information about the varieties in the DUS trial. The following information is necessary to control the data flow and to identify the data (Table 1):

Table 1: Key information in a DUS trial for winter wheat

<i>Information</i>	<i>Example</i>	<i>Comment</i>
crop	WW	winter wheat
year	2002	year 2002
location	1	Rethmar
characteristic	c007	length of plant
replication (plot)	r1	first replication
variety number	2605	number in the database
plot number	2	plot number in the field trial
plant number	20	20 th plant

3. All this information is necessary for constructing a field plan, for printing field labels for each plot and for other steps of the DUS test procedure. It is very useful to have special codes for the crops (e.g. abbreviation ‘WW’ stands for the crop winter wheat) and in certain cases for other information (e.g. trial station, name of characteristic, expression of characteristic, variety number and so on). The example used in the following case is for a trial sown at the trial station Rethmar, near Hanover in Germany.

4. Additional information is also included in the trial plan (Table 2). This additional information is not necessary to understand the structure of data but can be helpful to give more information.

Table 2: Additional information

<i>Information</i>	<i>Example</i>	<i>Comment</i>
group	1333	grouping information
applicant	SAATNORD	Company 'SAATNORD'
denomination	Diamant	proposed denomination
procedure	S	PBR-procedure
type	I	inbred line
sample	SM01	seed sample 2001

5. The trial plan includes all the information for more than 300 varieties for the year 2002. A part of this plan is included in Table 6. To save storage space, some information is not included in the trial plan on the datalogger.

III. Data capture and datalogger

6. It may be necessary, as a second step, to transfer some data from the personal computer to the program on the datalogger (for example, the field plan with the order of the varieties) in order to organise the recording process. The field plan on the datalogger with the order of the varieties in the trial is important for rationalisation of the data capture. This information avoids the need for an input of the variety number and thereby saves time during data capture. The data structure on the handheld computer is the same as on the PC. When using paper forms instead of a datalogger there is a requirement to transfer information from the field plan on the PC to a paper sheet manually or by using a special print program.

7. Different systems are used for storing data on the datalogger or on the PC. Two common systems are:

- storing data in files for each stage of evaluation
- firstly storing data in files and then transferring into a database.

8. It is necessary to store all the original data from single plants or from plots. For further computations, the storing of averages, variances and other statistical parameters is more practical than calculating these parameters several times.

9. To distinguish between '0' (zero) and ' ' (missing) it is important to use a definite code for 'missing values'. 'Zero' is generally not the method of choice for coding of 'missing values'. In particular for counts (which are a special class of characteristics) it is not appropriate to give the count 'zero' the code '0' and a missing count the same code '0'. A 'zero' count might be used, for example, for non-flowering plants. The cause of a missing value might be a faulty plant. Clearly, the same coding would lead to incorrect results (see column 2 in Table 3). Therefore a special code is needed for missing values.

Table 3: Example for handling of missing values

<i>Plant</i>	<i>Inflorescence: Number of flowers (incorrect)</i>	<i>Inflorescence: Number of flowers (correct)</i>
1	6	6
2	10	10
3	0	* (Missing value)
4	8	8
average	$(6+10+0+8)/4=6$	$(6+10+8)/3=8$

10. The correct result for the average number of flowers is 8 (see column 3 in Table 3). Depending on the software, special kind of codes are used e.g. blank (' '), dot ('.'), star ('*') or other. Using the figure "0" for missing values instead of a code may lead to a wrong calculation of the average.

11. The most efficient method for recording data is to use a datalogger, for example:

- Psion 'Workabout' (<http://www.pSION.com>)
- Husky 'FS4' (<http://www.wpihusky.com>)
- Infos 'PNT1800' (<http://infos-group.com>)

12. A docking station to charge the batteries and to transfer the data is also essential. Special software is required for the datalogger. There is not standard software available from the datalogger manufacturer and it will need to be developed by the user in cooperation with the manufacturer or with a specialised software company.

13. The advantages of dataloggers are:

- the elimination of hand-written field books
- the possibility of checking the range of data using predefined minimum and maximum
- the possibility of checking certain plausibility criteria (e.g.: characteristic A must always be greater than characteristic B)
- transfer of the field plan from the PC to the datalogger instead of using handwritten field books.

14. It is very useful to be able to download the prepared complete field plan from the PC to the datalogger with: plot number, variety number, replication number, ordered and grouped varieties, minimum and maximum values and plausibility criteria to be able to test the data during the input into the datalogger. For each crop, the datalogger needs a special set of characteristics.

15. To avoid errors, it is necessary to ensure that existing data cannot be overwritten accidentally. The use of a special "archives feature" for each characteristic is very helpful. If the archives feature is 'ON' it is impossible to overwrite the corresponding data. After running the archives procedure to transfer the data from the datalogger to the PC, the

“archives feature” will automatically be set to “OFF”. If the archives feature is “OFF”, the data can be overwritten or deleted on the datalogger.

IV. Data structure

16. Data must have a clear and stable structure. Certain variables are used as keys to define the structure. The most important key variables are crop, year, location, replication, variety number and plant number. Then a key must be established for the characteristics e.g. M1, M2, M3 and so on.

17. There is a requirement to differentiate between measurements (data with floating points) and visually assessed data (normally only a single digit). In the case of characteristics which have only one digit for the expressions, it is not necessary to complete the input procedure by pressing the “enter” key. The program is able to jump automatically to the next variety after the data input of the expression.

18. Handling of data for a single species is simple because the data structure is created for this species. However, if data for more than one species have to be handled, it is necessary to ensure that the names of characteristics are stored in a separate table and not with input data. The formal names of variables (characteristics) in the input data set are M1, M2, M3 and so on, for example (Table 4).

Table 4: Example for part of database with formal names of characteristics

<i>crop</i>	<i>year</i>	<i>location</i>	<i>replication</i>	<i>variety number</i>	<i>plant number</i>	<i>M1</i>	<i>M2</i>	...	<i>M13</i>	...
WW	2001	1	1	2605	1	2410	1711	...	108	...
WW	2001	1	1	2605	2			...	105	...
WW	2001	1	1	2605	3			...	110	...
WW	2001	1	1	2605	4			...	113	...
WW	2001	1	1	2605
WW	2001	1	1	2605
WW	2001	1	1	2605	18			...	105	...
WW	2001	1	1	2605	19			...	91	...
WW	2001	1	1	2605	20			...	102	...
WW	2001	1	2	2605	1

19. By using the key variable “crop”, it is possible to assign the name of the characteristic and other information from the separate table (Table 5).

Table 5: Example for a related “table” using the key variable “crop”

<i>crop</i>	<i>formal name of characteristic</i>	<i>internal database identification of characteristic</i>	<i>English name of characteristic</i>	<i>other</i>
WW	M1	b001	date of sowing	
WW	M2	b002	date of emergence	
WW	
WW	M13	b309	Ear: length cm	
WW		

This is a “relational database”.

20. In addition to the original input data some descriptive information about the characteristics should be stored, such as:

- name of characteristic in at least two languages (e.g. mother tongue and English)
- written states of expression in full (for the variety description)
- input and output format of data
- type of data (scale level) and formulae for calculation, if necessary (for automatic data processing).

21. The use of special evaluation procedures depends on scale level of data. Therefore for each characteristic, it is necessary to store this information in the database, in such a way that it will be possible to select automatically a set of characteristics for special evaluation procedures (for example: COY-U for measurements).

V. Database systems

22. There are many different types of database systems available. Different types of database systems are available depending on the hardware used. Examples of frequently used database systems are:

- Microsoft-Access for Windows-PC
- Microsoft-SQL-Server for Windows-PC
- Interbase for Windows-PC
- Oracle for several operating systems (Windows-PC and Unix machines)
- IBM/Informix or IBM/DB2 for several operating systems (Windows-PC and UNIX machines).

23. At present, it is not possible to use a database system with ideal performance for dataloggers. Databases are applicable for Windows-PC in the first step of storing the data and for central data management on central server systems (with a UNIX or Windows based operating system), but not on a datalogger with the operating system WindowsCE. All

calculation procedures need an interface between the program applied and the stored data. This interface may be:

- direct access to files
- direct access to a database (“native connection”)
- general access to a database using special interfaces (ODBC, JDBC,...).

24. If no database system is in use, data must be stored using direct access from program to file. The greatest advantage of using native connections is the speed of data transfer. The disadvantage is the requirement to rewrite this access program if the database is changed. “Native” means that the program only works with the corresponding database and not with any others. The best method is to store data in a database by using ODBC-interface or JDBC-interface for JAVA. This procedure guarantees independence from the database system and means that it is possible to change the database system without rewriting the application program.

VI. Data backup

25. A separate backup strategy should be established according to the hardware used. For dataloggers, the first rule is to download the input data daily onto the PC, i.e. after each working day in the field, the data should be transferred from the datalogger to the PC using a docking station. Depending on the type of datalogger, it may also be necessary to charge the batteries on a daily basis.

26. The second rule is never to store data on a local PC only. The best way to store data is by using a connection to a server where a central data management service is responsible for storing all data on external tapes. There are various procedures for storing data on tapes. For example, weekly storage on tapes of the whole set of data, whilst on the other days, only the data that have been changed since the day before (delta-backup) are copied and the differences stored on the tape.

27. If there is no central server for data management, the crop expert should organise a separate management procedure to store the data with an appropriate level of security. The storage procedure should be able to recover any data under all circumstances.

Table 6: Part of cultivation plan for the crop Winter Wheat in the year 2002 (location: Rethmar)

2002											
<i>serial number</i>	<i>crop</i>	<i>variety number</i>	<i>applicant</i>	<i>denomination</i>	<i>procedure</i>	<i>years</i>	<i>type</i>	<i>sample 1</i>	<i>sample 2</i>	<i>sample 3</i>	<i>comment</i>
			group:1333								
1	WW	6506	DESPREZ	Soissons	RES	2/-	I	PZ00			§55
			group:1335								
2	WW	2605	SAATNORD	Diamant	S/-	2/-	I	SM01			EUS
3	WW	1904	SEMUNDO	Tambor	-/L	-/9	I	SM98			EUS
4	WW	3012	JAHN-DES	JD-99-GZV 2001	1/-	3/-	I	RP00			
			group: 1337								
5	WW	3080	EGER	PBIS 00/77	-/1	-/2	I	W 02	W 01		
6	WW	3082	EGER	PBIS 00/91	-/1	-/2	I	W 02	W 01		DK1
			group: 1353								
7	WW	2905	STRUBE	Xenos	-/-	2/-	I	RP01	PZ02		EU1
8	WW	1641	LOCHOW	Bussard	S/L	12/12	I	SM96	VGL		
			group: 1355								
9	WW	6559	SERASEM	Tremie	RES	2/-	I	PZ01			
10	WW	2101	PETRSNFA	Brigadier	S/-	8/-	I	SM95			
11	WW	2471	EGER	Kris	-/L	-/3	I	SM00			
12	WW	2490	PETRSNFA	Maverick	-/L	-/3	I	SM00			
13	WW	2335	EGER	Kornett	-/L	-/5	I	SM98			
14	WW	3025	BENOIST	Ordeal	RES	2/-	I	PZ01			
15	WW	3028	ADVANTA	Savannah	RES	2/-	I	PZ01			
16	WW	1889	CEBECO	Ritmo	S/L	9/9	I	SM01	VGL		EUS
17	WW	2925	INTRSAAT	ISZ 59	-/1	-/3	I	W 01	W 02		
...											