Indian J. Pl. Genet. Resources 5(1): 57-66, 1992

COMPUTERIZED DATA PROCESSING AND CATALOGUING OF PLANT GENETIC RESOURCES

R.C. Agrawal and R.K. Mahajan

National Bureau of Plant Genetic Resources
New Delhi – 110012

Enormous information has accumulated over the past few decades on germplasm of various crops and their wild relatives. Awareness and importance of computerisation of database information, loging, analysis and retrieval is increasing day by day at national, regional and global levels. Various steps in preparation of a catalogue vis-a-vis recording of data, and duplication, verification, sorting, analysis, querying and listing of data are discussed in this paper. Computer program based on dBASE III Plus software for identification of duplicates as well as sorting and analysis of data are also discussed.

With tremendous pile-up of information by various global and national agencies, viz., CGIAR institutes, USDA, NBPGR etc., on the collection, characterization and evaluation, storage, *in situ* conservation and usage of the germplasm of various crops and their related wild species during the past few decades, scientists now-a-days spend at least 30 per cent of their time in taking care of the data information associated with collections (Rogers *et al.*, 1975). Their accumulation to the long-term genebanks at various stages of documentation posed serious threat to their efficient management as well as utilization by breeders. IBPGR has compiled directories of germplasm resources in various crops available with various national, regional or international organisations. In most cases, the stage of documentation is manually operated data registers or partially computerised data banks. Bettencourt and Konopka (1990) have indicated recent trends in documentation of sorghum germplasm database at 51 international centres. There is a clearcut advancement towards switching over to computerised database management. But since no standard data retrieval system

exists at different plant genetic resources organisations, except a few, e.g., GRIN of USDA, IRRIGEN of IRRI etc., the inter-institutional interaction among users becomes quite a difficult task. In the modern era, electronic computers are extensively used in different organisations for multiple purposes. Geographic Information Systems (GIS) as developed by Global Resources Information Database (GRDA) are powerful tools for integrating and analysing diverse spatial data (Burrill *et al.*, 1991). Data input, information and exchange through magnetic floppies, magnetic tapes, compact disks, etc. becomes easier through systematic use of computers. A simple and common method for exchange of such information is through crop catalogues which are the printed versions of systematically pre arranged/analysed data based on a particular set of descriptors and descriptor states decided for a crop or its related species.

Crop catalogues are valuable source of information on passport data, various agrobotanical and economic characters and other traits of interest to research workers involved in crop improvement programmes. In the present context, a crop catalogue may be considered primarily a systematic listing of different accessions of a crop, their identification source, characterisation and evaluation data alongwith additional/optional pre-analysed information using different statistical tools. Extensive data processing is involved in bringing out a catalogue. The computer data processing is a series of operations carried out in order to convert the data into useful information. The voluminous data recorded over a period of time, its tabulation etc. and lengthy calculations done manually has been a tedious task for huge pile-up of germplasm information and thus switching over to computer data processing is a good proposition. Following paragraphs relate to stepwise information required for preparation of a catalogue for efficient retrieval of information database.

CATALOGUING AND DATA PROCESSING

Cataloguing involves frequent use of many softwares and statistical packages, viz., dBASE III Plus, Fox Pro, Reflex, Lotus 1-2-3, MSTAT-C, SPSS, BMDP, MICROSTAT, Harward Graphics, Word Star, Word Perfect etc. Among these, the most frequently used package is any DBMS (Data Base Management System) package. A DBMS package is a computer program that organises large collections of interrelated data according to a well defined

scheme and produces reports by extracting and reformatting data. Reformatting refers to indexing and/or sorting on a particular field(s). It may further range from rearrangement or simple totalling to complex statistical analysis. Thus, a DBMS is a system for information storage, retrieval and analysis with emphasis on formatted data. In all the standard softwares on DBMS, we come across two terms — Field and Record. A Field can be considered as any descriptor [a widely accepted computer term for the character of plant, as well as for other units of information such as country of origin or the date of collection (Sapra and Singh, 1992)] from the list of descriptors whereas a Record is an accession with all descriptors. A collection of related records/ information constitute a file in computer environment.

Recording of data

The information about passport data (accession identifiers and information recorded by collectors) and characterisation (recording highly heritable characters capable of expressing themselves in all environments), preliminary and further evaluation is recorded on paper or on data loggers, a handy device, from which data can be subsequently transferred to a computer for further storage and analysis. To facilitate description of accessions, results of tests of accessions and to improve communication between scientists in different institutions, internationally accepted IBPGR descriptors lists decided for each crop species should be followed. These norms are well defined for qualitative as well as quantitative characters (Rogers *et al.*, 1975). For scoring in heterogeneous populations Hintum (1989) approach should be followed.

Various statistical designs are available in literature which can be extensively used depending upon the population size, availability of seed and resources. In plant genetic resources, the accessions for different crops vary invariably from hundreds to thousands. Limited seed quantity in each accession and meagre experimental land makes it difficult to replicate the accessions and thus the basic norms of statistical designs are violated. In such situations, augmented randomized design may be adopted (Federer, 1956). The design consists of one set of treatments (the check varieties) replicated a number of times and a second set of treatments (the accessions) appearing only once in the entire layout. A computer program for statistical analysis of augmented randomized complete block design was further described in simplified form (Sapra and Agrawal, 1992).

Loading of data to computer

The data is transferred to the computer system generally by data entry operators. It is normally carried out either in text mode by using some text editors, *viz*, Norton Editor, Word Star (Non-document mode), Edit etc. or by directly using the DBMS package. Spreadsheets like LOTUS 1-2-3, Quatro etc. may also be used for this purpose. For fast data feeding, data entry machines may also be used. If data is stored on a data logger, the above steps can be skipped. These loggers minimise human error in data feeding. In text editors, the fields should be preferably separated by comma (,) which makes easy transfer of data to other software. In any DBMS package, extra care is taken while selecting the field width (the maximum width for different fields). Furthermore, a range may be defined for every numeric field so that the values outside the range will not be accepted at data feeding time.

Verification of data

For verification and subsequent use, the data loaded in the text editor should be transferred to DBMS package (preferably dBASE III Plus). The technique of transfer differs from package to package. For dBASE III Plus, the structure of fields should be defined first and then for transfer APPEND FROM command may be used. The error occurs in data recording and data feeding. Duplicate entries detected first and after manual checking from the original, such record may be deleted. Equal field width should be maintained for each accession throughout. For example two accessions, *viz.*, IC-532 and IC-1000 should be entered as IC-0532 and IC-1000 fixing a maximum field width of 7 characters. For identifying the duplicate entries, following program in dBASE III Plus can be utilised.

```
SET TALK OFF
SET EXACT ON
   USE filename-1
      INDEX ON fieldname-1 TO filename-2
      COPY STRUCTURE TO filename-3 FIELDS fieldname-1
         SELECT A
            USE filename-1 INDEX filename-2
         SELECT B
            USE filename-3
         SELECT A
```

```
        GO TOP
        AN=fieldname-1
        SKIP
    DO WHILE .NOT. EOF ( )
        IF fieldname-1=AN
    SELECT B
        APPEND BLANK
            REPLACE fieldname1 WITH AN
                ELSE
                    AN=fieldname1
                    ENDIF
    SELECT A
        SKIP
  ENDDO
CLOSE ALL
SET EXACT OFF
SET TALK OFF
```

In this program, characters in italics can be replaced by the user according to the choice.

Characters in italics	Explanation
Filename1	datafile
Filename2	file to store the indexed information
Filename3	file to store duplicates
Fieldname1	file name defined for accessions

For major data feeding errors, indexing on different fields may be carried out which can help in identifying the maximum and minimum values of different fields. For coded fields like seed colour, pod type etc. 'UNIQUE' indexing may be done. In case some wrong entry is detected, the entire data of the record should be varified.

Example-Following data has been extracted from a catalogue on Forage Sorghum Germplasm (Mathur *et al.*, 1992) for verification of duplicates. One accession at the last has been added as a duplicate accession. This catalogue contains 3943 accessions, and thus makes physical verification of duplicates a difficult task. Field names used to store 5 descriptors in dBASE III Plus are mentioned in the following column headings (Acc _no is for accession no., P_height is for plant height and S_thick is for stem thickness).

Acc_no	Pedigree	Location	P_height	S_thick
IS-01002	Gund	Maharashtra	187.6	16.0
IS-01015	—	Haryana	200.03	21.0
IS-01018	IC-3666	Haryana	193.3	21.0
IS-01023	Delhi Local	Delhi	228.3	16.0
IS-01029	Ramkel	Maharashtra	271.3	31.0
IS-01036	Palitina Dholiao	Gujrat	109.3	31.0
IS-01047	CO18	Tamil Nadu	192.0	16.0
IS-01113	Baligaon	Bihar	199.0	31.0
IS-01015	Bhindi Loha	Maharashtra	130.0	16.0

The file names and field names used are –

1. Sorghum. dbf as filename1
2. Accession. ndx as filename2
3. Dupli. dbf as filename3
4. Acc_no as fieldname1

After replacing these file names and field names in the program of duplicates, it can be run and a list of duplicates can be taken from the database file Dupli. dbf (for the present example, only one duplicate i.e. IC-01015 will be stored in this file).

Sorting and analysis of data

Depending upon the requirement, one can disturb the sequence of data records by sorting them on a particular field(s), say arranging the data records on the basis of increasing order of accessions. Before the data is subjected to sorting and statistical analysis, it is of utmost importance to have a duplicate listing of original data on some storage device so that the original data can be further used in future or in accidental damage while processing the data. Depending upon the number of descriptors, in case of larger field length it may be separated into two parts-passport data and characterization and preliminary evaluation data.

To gather some more important information, the data should be subjected to statistical analysis using standard packages, *viz.*, MSTAT-C, Lotus 1-2-3, Statgraf, Spar 1, Harward Graphics, Microstat etc. These packages can be utilised for presenting the information in histograms, bar diagrams, pie charts, regression analysis, correlation analysis, measures of central tendency and dispersion (mean, range, skewness, kurtosis, coefficient of variation

etc.), path analysis, stability analysis etc. Statistical analysis will enable the user in querying the database on individual and interrelated descriptors for selecting the desirable accessions for utilisation in crop improvement programs. For augmented design, the adjusted treatment means may be obtained by using Augment 1 package (Sapra and Agrawal, 1992) and may be incorporated in the data file as a separate field. These adjusted means should be used for further retrieval and querying purposes. The outcome of the analysis is presented in a suitable form by adding table headings, legends etc. Such modification is possible by using some text editor(s).

For computing the frequency distribution of qualitative descriptors e.g. country, state, district, colour, leaf type etc. the following dBASE III Plus program may be utilised.

```
SET TALK OFF
  USE filename1
    INDEX ON fieldname1 TO filename2
    SELECT A
      USE filename1 INDEX filename2
    SELECT B
      USE filename3
      ZAP
    SELECT A
      GO TOP
      temp1=fieldname1
      SKIP
      FRQ=1
    DO WHILE .NOT. EOF ( )
      IF fieldname1=temp1
        FRQ=FRQ+1
      ELSE
        temp2=fieldname1
        SELECT B
        APPEND BLANK
        REPLACE fieldname2 WITH temp1, fieldname3 WITH FRQ
        temp1=temp2
          FRQ=1
      SELECT A
    ENDIF
  SKIP
ENDDO
```

```
SELECT B
   APPEND BLANK
      REPLACE fieldname2 WITH temp1, fieldname3 WITH FRQ
CLOSE ALL
SET TALK ON
```

In this program, characters in italics can be replaced by the user according to the choice.

Characters in italics	Explanation
Filename1	data file
Filename2	file to store indexed information on fieldname1
Filename3	file to store frequencies (must be created before using the program)
Fieldname1	field name for which frequencies are to be computed (from filename1)
Fieldname2	field name in filename2 defining the field corresponding to fieldname1 of filename1
Fieldname3	field name in filename2 to store the frequencies

Example — In the earlier mentioned data (after removing the last record which is a duplicate accession), frequencies of accessions according to location can be calculated by replacing the following file and field names in the above mentioned program.

Sorghum. dbf as filename1
Location. ndx as filename2
Freq. dbf as filename3
Location as fieldname1
Location as fieldname2
Frequency as fieldname3

After the program is executed, file Freq. dbf will contain

Location	Frequency
Bihar	1
Delhi	1
Gujarat	1
Haryana	2
Maharashtra	2
Tamil Nadu	1

Querying and listing of data

To enable the user of a catalogue to select some promising accessions for important traits, some sample queries may be carried out using DBMS package and can be presented in a tabular form. For example, one can take list of accessions for forage sorghum, satisfying a query e.g. all accessions with days to 50 per cent flowering between 60 to 80 days, stem thickness less than 3 cm, number of leaves greater than 15 and forage yield greater than 2 Kg. (Mathur *et al.*, 1991). The list will include :

Accession number	Days to flowering	Stem thickness	Number of leaves	Fodder yield
IS-27087	72.0	2.83	15.2	2.2

A hard copy of the analysis part, querying part and passport and evaluation data may be taken by using some printing device. The duplication of print outs may be taken on offset printing, xerox machines etc.

The discussion above is extremely helpful in bringing out salient features of data and also systematically preparing a catalogue. Moreover, one can improve upon as per requirements and availability of resources.

ACKNOWLEDGEMENTS

The authors are grateful to Dr. R.S. Rana, Director, NBPGR, New Delhi, for the facilities provided.

REFERENCES

Bettencourt, E. and J. Konopka. 1990. *Directory of germplasm collections.* Vol III, Cereals. Rome, Italy, IBPGR

Burrill, A., H. Cronze and O. Simonett. 1991. The potential use of the global resources information database (GRID) in plant genetic resources activities. *In: Crop genetic resource of Africa.* Attere, F., H. Zedan., N.Q. Ng., P. Perrino (eds.). Vol. I. Proceedings of an international symposium, Nairobi, Kenya, 26-30 September 1988 Rome, Italy, IBPGR: 125-132

Federer, T. Walter. 1956. Augmented (or Hoonuiaku) designs. The *hawaiian planter records.* Vol IV, 2nd Issue, 1956: 191-208

Mathur, P.N., K.E. Prasada Rao, T.A. Thomos, M.H. Mengesha, R.L. Sapra and R.S. Rana. 1991. *Evaluation of Forage Sorghum Germplasm : Part-I.* NBPGR Pusa Campus, New Delhi

Mathur, P.N., K.E. Prasada Rao, I.P. Singh, R.C. Agrawal, M.H. Mengesha and R.S. Rana. 1992. *Evaluation of Forage Sorghum Germplasm: Part-II.* NBPGR, Pusa Campus, New Delhi: 296p

Rogers, D.J., B. Snoad and L. Seidewitz. 1975. Documentation for genetic resources centres. *In: Crop genetic resources for today and tomorrow.* Frankel, O.H. and J.G. Hawkes (eds.). IBP 2, Cambridge University Press, Cambridge: 399-405

Sapra, R.L. and R.C. Agrawal. 1992. Germplasm Evaluation: Augmented Designs. *In: Plant genetic resources: Documentation and information management.* Rana, R.S., R.L. Sapra, R.C. Agrawal and R. Gambhir (eds.), NBPGR, New Delhi-12: 37-43

Sapra R.L. and Bhag Singh. 1992. Database management of plant genetic resources *In Plant genetic resources: Documentation and information management.* Rana, R.S., R.L. Sapra, R.C. Agrawal and R. Gambhir (eds.), NBPGR, New Delhi-12: 17-35

Hintum, Th. Van. 1989. Scoring hetrogeneous populations. *In:* Report of an international workshop Beta genetic resources. IBPGR publication. International crop network, seriese 3: 106