Towards an effective collection, organisation and storage of data in agroforestry research

The increasing use of systems analysis in current agroforestry research highlights the need for a standardised procedure for data collection and management. An easier way to satisfy this demand could be a development of the ICASA V 1.0 Data Standard, already extensively used for a rational storage of data and metadata in agricultural science. However, as its structure was formulated to feed FORTRAN-DSSAT models, designed mainly for tropical and subtropical crops, it is difficult to use in agroforestry where the researchers deal with complex interactions between trees and crops with multi-dimensional levels of measurements.

Introduction

The use of systems analysis in several branches of applied science from ecology to soil science has increased tremendously in the last 10 years due to the widespread use of personal computers and modelling packages and to the demand of 'IF THEN' scenarios from policy makers and other stakeholders (Bouma et al. 1996). This approach requires the design and validation of biophysical models with different levels of complexity, as well as the availability of large and organised sets of experimental data for their evaluation and validation (Hunt and Boote, 1998). The dataset must describe a whole experimental environment with as many parameters as are necessary to describe it as a system. The data must be easily accessible to the whole interested scientific community, not only the current community but future members also. Real sharing implies also that the set can be edited and updated by workers not at all involved in its original making. This means that all the information on how the data were measured and treated (transformed or extrapolated) must be also made available as well all the instructions needed to understand them.

Therefore the collection and the storing of the data produced during a scientific investigation, which is and always has been one of the main tasks carried out by scientists, is becoming a critical step in any scientific project, demanding systematic and standardised techniques (Michner et al., 1997; Boone et al., 1999).

Data and metadata

The literature on large datasets agrees that supporting documentation must be exhaustive to interpret a dataset, detailing methods, variables and units (Boole et al., 1999). This documentation is called "metadata", or literally "what is beyond the data". Boole et al. (1999) stated that the omission of this information from a set of scientific data can "strongly limit the use of the data beyond the original study". The collection of metadata has always been done concurrently with the collection of experimental data, but rarely has it been done systematically. It is the lack of standards in metadata collection which generates the "information erosion" (Figure 1) more than the complete absence of the metadata.

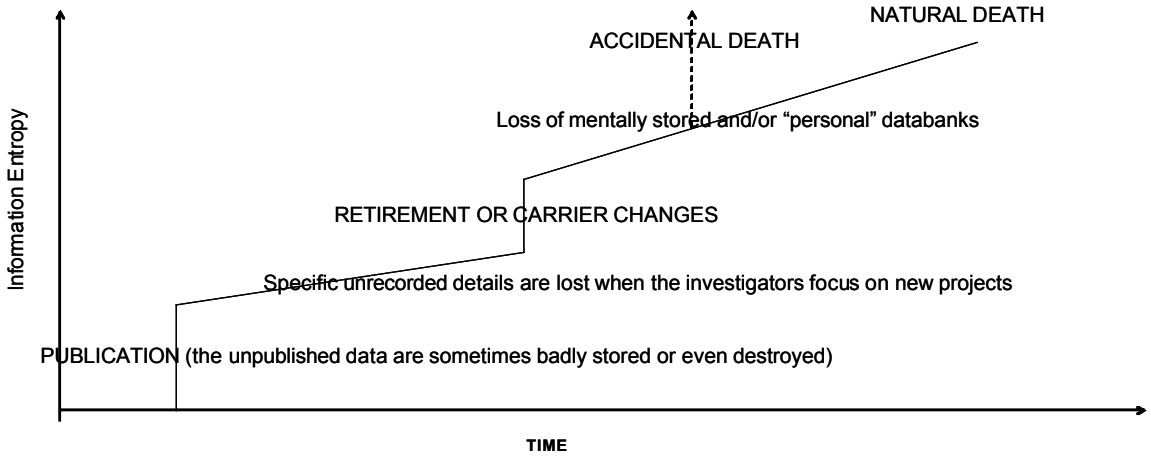

Figure 1. The quality of information stored in a set of experimental data degenerates with time. The storage of "on the side" information relies mainly on informal databanks belonging to the personnel involved in the research, which ultimate follow their human fate (adapted from Michener et al. 1997).

The creation of standards is nothing more than a sharing of mutually agreed conventions. In the case of collection and storage of information, common conventions are needed in at least four areas (Uehara and Tsuji, 1993): the codes for variables and measures, the vocabulary for documents, the structure of the data file, and the minimum set of

data required to allow a model to simulate the experimental system.

Relational and non-relational databases

The simple editing of data in digital form in a file, however standardised its structure, is not sufficient for easy retrieval by someone who did not set up the database initially. Setting up a database must be done in such a way that editing and updating, as well as retrieving of data across different sets, requires a minimal number of operations. The development of the appropriate techniques for this purpose constitutes most of a specialised branch of information technology (Carte, 1995): Data Base Design (DBD). If a large set of data is organised according to the criteria of DBD, we are dealing with a so called "relational database", which through links between data grouped in different objects (tables), allows faster and easier management and retrieval and use of data and information. A relational database can be "queried", extracting part of the data without dealing with all the stored information, and can be modified by editing only those records you require to change.

Effective storage of experimental data for systems modelling does not require a relational database, however if, together with the above mentioned conventions, a set of data is organised in digital form using the techniques of DBD, its subsequent use in modelling exercises will be greatly facilitated.

Development of standards in agricultural research: IBSNAT and ICASA

Agroforestry is a specialised branch of agricultural science, which studies the dynamic interactions between trees and crops in a single system. However, the research environment and the data produced can be quite different in detail and quality from those obtained from experimental plots with monocultures e.g. a sole crop or trees in woodland or forest. So far no specific standard has been developed to handle research data from agroforestry plots. To ease this task, the starting point could be the state of the art on advanced data storage in agricultural science. This would give an existing tool for archiving crop data, which could be improved and redesigned to allow the management of data from the mixed crop (arable or pasture) and tree plot designs of agroforestry.

IBNSAT

A first attempt to set an internationally accepted standard for recording data in agricultural science was made by the International Benchmark Sites Network for Agrotechnology Transfer or IBSNAT (Uehara and Tsuji, 1993).

The need to have the same procedure of data management in different experimental sites managed by different research teams working together highlighted the necessity for standards for data and metadata storage. Therefore an agreed minimum set of data was developed (MSD) that was required to run a decision support model for a given crop. This set contained information on the experimental site, the weather, the soil, and the management and performance of the crop. In the case of surrogate (i.e. simulated or predicted) values, the procedures to obtain them were standardised. The data were exchanged between scientists in 25 countries through files which shared a common set of codes and a common structure. The rules to edit and modify the data files were also standardised (Hunt and Boote, 1998). The standards so developed were good enough to be used even in non-agricultural research, and were adopted by the Global Change and Terrestrial Ecosystem Project (GCTE).

However, this first exercise had several limitations. The file structure was determined by the need to feed a particular model (DSSAT), the codes were far from being self-explanatory, being written to be read by the Fortran language (hence limited to eight alphanumeric characters), and the files required a specific database to enable them to be edited (DBF). A further limitation to more extensive use was the fact that the IBSNAT project was focused on studying and simulating tropical crops.

Despite these limitations, a useful conclusion was reached, which was and is a basis for further development: the agreement on a minimal set of data - variables, constants and parameters, easy to be measured, and with enough detail to allow a computer simulation of several agricultural systems.

ICASA

The development of standards for data management in agricultural science has not stopped with the IBSNAT project but has become part of the objectives of the International Consortium for the Agricultural Systems Application (ICASA) which currently pursues some of the research programs started by IBSNAT (Bouma and Jones, 2001). The structure and the codes already designed within the IBSNAT consortium were further developed and the current standard, known as the 'ICASA V 1.0 Data Standard', defines data files as ASCII text files with 254 characters per line so that each file can be edited by a simple text editor and analysed by any spreadsheet software. However the IBSNAT structure has been kept to allow the direct input of data to DSSAT models, consequently the limitation of non-self-explanatory codes remains.

ICASA standard - advantages and limitations

The typical structure of an "ICASA" file organises the data, with increasing level of aggregation, in lines of codes (beginning with @) each called a "data cluster", more clusters describing the same event or properties are joined in groups, i.e. Treatments, Cultivars, Fertiliser. Eventually more groups form a Data Unit (beginning with $), four standard data units have been so far defined by ICASA: Experimental Details, Plot data, Soils and Weather. The Units can be edited with each in a single file or in a combined experimental file. The single data items are linked between different sections of the Units by a Level Indicator or "key",

which has a similar function to the" primary key "and a "foreign key" in a relational database.

A further important convention established by ICASA deals with the file names, in which are summarised information on what kind of data are in the file, from what experiment, carried by who, when and where, and with what kind of crop. Each file name is composed of a prefix of eight characters and an extension of three characters. In the prefix there are codes of two characters each for the institution carrying out the experiment, for the site, the year, and the experiment number. The extension indicates the kind of Units (Experimental details: X, Plot data: D, Soils: SOL, Weather: WTH) present in the file and, for type 'Experimental', the crop involved (cc). Therefore the file name ULCR9902.WWX indicates the file with experimental data (X) for winter wheat (WW) acquired by partners at Cirencester (CR) in 1999 (99) for the second silvoarable experiment (02) of the University of Leeds (UL).

The file structure, proposed in the V 1.0 ICASA Data Standard (Hunt et al., 2001), allows easy transfer and editing of the data, but is far from being realisable as a relational database, even if assembling data in clusters and groups simulates a 'table-objects' style of organisation and the Level code allows links among different clusters similar to those created by primary and foreign keys as in a real relational database. The files must be interpreted by the researcher using a glossary of codes and cannot be handled directly by database software.

However the standard is already a powerful tool, tested by several research groups in different parts of the world and under continuous improvement. Its use, thanks to a very general approach, is already not limited to pure agricultural research. It not only allows rational storage of the data based on common rules, but also the recording of any metadata as a string, in a defined section of a text file. The end result combines data in a simple spreadsheet, (e.g. see Appendix Table 1) ready to run simulation models with the metadata required for their understanding in a single text file with a conventional set of explanatory information (e.g. see Appendix Tables 2a, 2b and 2c). The data organised as a set in Appendix Table 1 cannot be shared outside the original research team, the same data set organised according to the ICASA standard in Tables 2a, 2b and 2c, can be read easily by anybody with publicly available knowledge of the use of the ICASA codes and rules (http://www.icasanet.org/).

Possible improvement and the specific requirements of agroforestry research.

The use of a data standard in agroforestry requires, of course, the development of specific codes for measures of trees, and for describing the agroforestry plot with its spatial interaction between crop and tree at root and canopy level.

The ICASA standard, the current state of the art for data storage in agricultural science, has not so far created such codes and its general approach, aiming to limit as much as possible the proliferation of code items, makes it difficult to create new code for a limited research field such as agroforestry.

Further demand, not only in agroforestry but in any research field, is for a data management tool, which, without losing the advantage of a standard based on text files, allows, however, an organisation of data as close as possible to that done by a relational database. This demand has been partially satisfied by the development of an ACCESS database able to be fed directly by ICASA text files, by researchers at the North Carolina State University, a group already involved in the ICASA consortium. However a more transportable tool which can run in all operating systems and can be edited "on line" would be preferable.

Finally the codification limited to eight digits should be abandoned and the use of data files within more up-to-date models, created by current programming languages and techniques should be contemplated.

The future work for researchers involved in agroforestry experiments will be exploiting and improving the standard proposed within the ICASA, persuading this institution and the larger agricultural science community of the special needs of researchers dealing with data from tree-crops. On the other hand, system analysts must include in the design and development of up-to-date models of agroforestry systems databases that are easily transportable and accessible on the WWW network using Standard Query programming Languages (SQL) as has already been done in other fields.

F. Agostini, D.J. Pilbeam and L.D. Incoll
Faculty of Biological Sciences, L.C. Miall Building, University of Leeds, Leeds LS2 9JT, UK.

E-mail: f.agostini@leeds.ac.uk.

Acknowledgements

One of us (FA) was funded by the European Union as part of the Silvoarable Agroforestry for Europe (SAFE) project.

References

Boone, R., Grigal, D.F., Sollins, P., Ahrens, R.J., Armstrong, D.E. (1999) Soil sampling, preparation, archiving and quality control. In: Roberston G.P., Coleman D.C., Bledsoe C., Sollins P. Standard Soil Methods for Long-Term Ecological Research. Oxford Press University, Oxford, UK, pp. 21-28.

Bouma, J.and Jones, J.W. (2001) An international collaborative network for agricultural systems applications (ICASA). Agricultural Systems 70, 355-368.

Bouma J, Van Keulen H., Van Laar, H., Rabbinge, R. (1996) The 'School of De Witt' crop growth simulation models: A pedigree and historical overview. Agricultural Systems 52, 171-198.

Carte, J. (1995) The relational databases. Chapman & Hall, London Hunt, L.A. and Boote, K.J. (1998)

Data for model operation, calibration, and evaluation. In: Tsuji, G.Y. Hoogenboom, G. and Thornton, P.K. Understanding Options for Agricultural Production, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 9 -39.

Hunt, L.A., White, J.W. and Hoogenboom, G. (2001) Agronomic data: advances in documentation and protocols for exchange and use. Agricultural Systems 70, 477-492.

Michener, W.K. Brunt, J.W., Helly, J.J., Kirchener, T.B. and Stafford, S.G. (1997) Nongeospatial metadata for the ecological sciences. Ecological Applications, 7 (1), 1997, pp. 330-342.

Uehara, G. and Tsuji, G.Y. (1993) The IBSNAT project. In: Penning de Vries, F.W.T. Teng, P.S Metselaar, K. Systems Approaches for Agricultural Development. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 505 -513.

APPENDIX

Table 1. Effect of two levels of nitrogen fertilisation on a Maize/Barley rotation in the Padana Valley (Modena) (data of F. Agostini). The crop data plus soil and weather data in this table are sufficient to run three different crop models (SUNDIAL, CROPSYS, NCSOIL). However the table is not self explanatory and it does not deliver enough information to allow any modeller anywhere, other than the experimenters, to run the models. na = not available.

Treat_NAME	Year	Crop	Prev-crop	Prev_yie	Prv_Harv_wk	Year_Start	Sow_wk	Harv_wk	Yield	N_in(kg/ha)	N-inAppl_wk	L_	L_appli_wk
Control	94	BA	MZ	7.17	2/9/1993	3	21/10/93	4/6/1994	2.20	0.00	0.00	0.00	0.00
M1	94	BA	MZ	8.25	2/9/1993	3	21/10/93	4/6/1994	6.43	100	18/03/94		
M2	94	BA	MZ	8.25	2/9/1993	3	21/10/93	4/6/1994	7.47	200	18/03/94		
L1	94	BA	MZ	8.68	2/9/1993	3	21/10/93	4/6/1994	3.28			20.00	13/10/93
L1M1	94	BA	MZ	8.16	2/9/1993	3	21/10/93	4/6/1994	6.61	100	18/03/94	20.00	13/10/93
L1M2	94	BA	MZ	7.60	2/9/1993	3	21/10/93	4/6/1994	7.21	200	18/03/94	20.00	13/10/93
Test	95	BA	MZ	4.72	2/9/1994	2	25/10/94	3/6/1995	3.42				
M1	95	BA	MZ	7.50	2/9/1994	2	25/10/94	3/6/1995	5.62	100	25/03/95		
M2	95	BA	MZ	8.25	2/9/1994	2	25/10/94	3/6/1995	5.81	200	25/03/95		
L1	95	BA	MZ	6.38	2/9/1994	2	25/10/94	3/6/1995	4.19			20.00	15/10/94
L1M1	95	BA	MZ	na	2/9/1994	2	na	na	na				
L1M2	95	BA	MZ	8.63	2/9/1994	2	25/10/94	3/6/1995	5.77	200	25/03/95	20.00	15/10/94

Table 2. Data and metadata for the experiment in Table 1 on the effect of two levels of nitrogen fertilisation on a Maize/Barley rotation in the Padana Valley (Modena) organised according to the ICASA standard. With knowledge of the ICASA standard the reader should be able to understand the experiment and its results.

Table 2a. Effect of two levels of nitrogen fertiliser on a Maize/Barley rotation in the Padana Valley (Modena) (data of F. Agostini). Metadata organised according to the ICASA standard, giving explanations of the data measurements and the experimental design

```
 EXPERIMENT: CADISA001
*GENERAL
@ NAME
  Effect of 5 fertilization regimes ( two levels of inorganic and 1 level of organic)
  on soil nitrogen (as nitrate) content under winter barley within a maize/winter barley rotation on two different fields.
@MAIN_FACTOR    FACTORS LOCAL _NAME
Nitrogen        2FE*1RE Cadriano
@ PEOPLE
  Dr. Marcello Donatelli, ISCI Bologna, Italy
@ VERSION
 01-05-02 F. Agostini (School of Biology, Leeds University).
@ OBJECTIVES
  To demonstrate the ICASA standard to the Agroforestry Research Group of the School of Biology. First simplified exercise.
@ MEASUREMENTS
  Nitrogen content, soil water content, yield of the crop.
@ METHODS
  Triplicate core sampling with an auger randomly at different depths, each 7 -15 weeks
  Nitrogen measured by Orion electrode.
@ PROBLEMS
  Four sampling operators involved.
@ NOTES
 It is just an exercise. A list with the ICASA item code used has been added at the end of the file to make it easier to read. The initial soil
nitrate content has been expressed in µg N/g soil although the equivalent ICASA unit is g N/Mg soil.
@ QUALITY
  The Orion electrode gives readings with an accuracy of 2 ppm, no error term available.
@ PUBLICATION
  Gabrielle R, Agostini F., Donatelli M. (2000) Limits of accuracy of the water component of a Decision-Support-Oriented Agronomic
Model, Ital. J. Agron. 3,2,87-99.
@ DISTRIBUTION
  As example of an ICASA standard file, only within the Agroforestry Research Group of School of Biology
! The data set misses management data, the soil data available are soil water content
! and soil inorganic nitrogen reported in the file group TIME_COURSE(SOIL) and
!  INITIAL_CONDITIONS. The weather data of year 1994 and 1995 can be supplied by the author.
```

Table 2b. Effect of two levels of nitrogen fertiliser on a Maize /Barley rotation in the Padana Valley (Modena) (data of F.Agostini). Information on treatments and crop production organised according to the ICASA standard.

@TRNO	FL	Cr	DAYR	MDAT	HWAM
1	1	MZ	1993	245	7.17
2	1	MZ	1993	245	8.25
3	1	MZ	1993	245	8.25
4	1	MZ	1993	245	8.68
5	1	MZ	1993	245	8.16
6	1	MZ	1993	245	7.6
1	2	MZ	1994	245	4.72
2	2	MZ	1994	245	7.5
3	2	MZ	1994	245	8.25
4	2	MZ	1994	245	6.38
5	2	MZ	1994	245	-99
6	2	MZ	1994	245	8.63
1	1	BA	1994	155	2.2
2	1	BA	1994	155	6.43
3	1	BA	1994	155	7.47
4	1	BA	1994	155	3.28
5	1	BA	1994	155	6.61
6	1	BA	1994	155	7.21
1	2	BA	1995	153	3.42
2	2	BA	1995	153	5.62
3	2	BA	1995	153	5.81
4	2	BA	1995	153	4.19
5	2	BA	1995	153	-99
6	2	BA	1995	153	5.77

*TREATMENTS			--			----------FACTOR LEVELS------- ----------'				
@TRNO	C#	O#	TREATMENT_NAME			CU	FL	SA	IC	PL	FE
1	2	1	CONT	1	1	1	1	1	0	0	1
1	2	1	CONT	1	2	1	-99	2	0	0	2
2	2	1	M1	1	1	1	2	1	1	0	1
2	2	1	M1	1	2	1	-99	2	1	0	2
3	2	1	M2	1	1	1	3	1	2	0	1
3	2	1	M2	1	2	1	-99	2	2	0	2
4	2	1	L1	1	1	1	4	1	0	1	1
4	2	1	L1	1	2	1	-99	2	0	1	2
5	2	1	L1M1	1	1	1	5	1	1	1	1
5	2	1	L1M1	1	2	1	-99	2	1	1	2
6	2	1	L1M2	1	1	1	6	1	2	1	1
6	2	1	L1M2	1	2	1	-99	2	2	1	2

Table 2c. Effect of two levels of nitrogen fertiliser on a Maize/Barley rotation in the Padana Valley (Modena) (data of F. Agostini). Information on crop management organised according to the ICASA standard.

*CULTIVARS

@CU	CR	CULTIVAR_NAME
1	BA	-99

*PLANTING

@PL	PL_NAME
1	1994_planting
2	1995_planting

@PL	PLYR	PLDAY	PLDS	PLDP
1	1993	294	PLD0B	5
2	1994	298	PLD0B	5

*FERTILISERS (INORGANIC)

@FE	FE_NAME			
1	Low	Input	at	spring
2	High	Input	at	spring

@FE	FEYR	FEDAY	FEDEP	FEAMN	FEAMP	FEAMK	FEAMC	FEAMO	FEOCD
1	1994	77	5	100	-99	-99	-99	-99	-99
1	1995	84	5	100	-99	-99	-99	-99	-99
2	1994	77	5	200	-99	-99	-99	-99	-99
2	1995	84	5	200	-99	-99	-99	-99	-99

*RESIDUES AND ORGANIC FERTILISERS

@RE	REYR	REDAY	RESTG	RECD	REACD	REDEP	REINP	REAMT	RESN	RESP	RESK
1	1993	286	-99	RE003	REA02	40	-99	20000	5	-99	-99
1	1994	288	-99	RE003	REA02	40	-99	20000	5	-99	-99