ClimagriLT: a Relational Meta-Database for Data Management of Long-Term Agronomic Experiments

M. ZULIANI, A. PERESSOTTI, G. ZERBI, G. ZULIANI, G. DELLE VEDOVE, and F. DANUSO

Dipartimento di Scienze Agrarie e Ambientali, Università di Udine, Italy

Corresponding author: G. Zerbi, Dipartimento di Scienze Agrarie e Ambientali, via delle Scienze 208, 33100 Udine, Italy. Tel.: +39 0432 558618, Fax: +39 0432 558603, E-mail: zerbi@uniud.it

Received: 10 June 2003. Accepted: 4 December 2003.

ABSTRACT

BACKGROUND. Data obtained from long-term agronomic experiments are a basic starting point for the development, calibration and validation of mathematical models of cropping systems. The usefulness of these experiments done in Italy is strongly impaired by scarce information on easily available data, on their organization and data collecting methods, and by difficulty in accessing them.

METHODS. The philosophy and layout of a meta-database (ClimagriLT) designed to store and share data related to a set of long-term experiments (treatments, yields, soil and meteorological data etc.) is presented in this paper. The main focus is on data model and management policy of data distribution. The details of the long-term experiments, i.e. locations, treatments, studied factors, data collection methods, availability, distribution, experiment site map are contained in the meta-database. The database stores weather data, soil analysis data, yield and biometrics of crops. The data model was built following the relational database theory because it is widely accepted, intuitive and easy to implement. Four steps drove the data model construction: i) requirements; ii) draft of the conceptual structure; iii) draft of the logical structure; iv) normalization. A brief introduction is also given to future applications.

Key-words: data model, shared data, metadata definition, time series.

INTRODUCTION

Studies on agro-ecosystems are a long-term challenge just as agriculture is a long-term enterprise. Shifting the time perspective from short to long-term is essential to understand the dynamics of these managed ecosystems which, as a consequence of increasing land exploitation and human pressure, cover a relevant fraction of the planet (Army et al., 1991).

Traditional agronomic experiments and empirical knowledge of agro-ecosystems are not sufficient to fill the complex matrix of indicators that decision-makers require to manage the overall agricultural system at regional and farm level (Jones et al., 2003). It is widely accepted that a possible approach to shift the time scale from short to long-term and improve knowledge on agro-ecosystems is through the use of mathematical models. Data obtained from long-term agronomic experiments are a basic starting point for the development, calibration and validation of reliable models aimed at improving their prediction capability and detecting lacks in knowledge and research activity (Acock and Acock, 1991).

The usefulness of long-term agronomic experiments is strongly impaired by scarce information on their organization and data collecting methods. Another obstacle to the full potential of their value is the lack of a common standard for data storage and management (Hunt, 1998; Hunt and Boote, 1998). Long-term experiments frequently pose a series of problems deriving from inconsistencies due to changes in experimental layout, transfer of responsibility from one scientist to another, reductions in necessary funding, etc. Collecting the complete documentation of a single long-term experiment is often difficult: loss of data, changes in analysis methods, etc., create problems with experiment reliability and data protection. The outcome is that data exchange among scientists and full information retrieval from experiments are not frequent, while errors are common.

A number of scientists involved in projects in different research fields (plant breeding, modelling, agro-meteorology, soil science, plant

physiology) recently began to work on this topic to find solutions for more efficient data storage and management (Van Evert et al., 1999a; Hunt et al., 2001). The relational database theory (Codd, 1970) has been widely adopted mainly because it represents a solid theoretical approach to the problem. Hierarchical databases and standardized text files have also been tested. The latter method was used by IBSNAT (International Benchmark Sites Network for Agrotechnology Transfer) to develop a set of text files that were recently upgraded by the International Consortium for Agricultural System Application (Hunt et al., 2001).

The philosophy and layout of a meta-database designed to store and share data related to a set of long-term experiments (treatments, yields, soil and meteorological data, etc.) is presented in this paper. The focus is on the data model and management policy of data distribution. A brief introduction to future applications is also given.

Long-term agronomic experiments in Italy

A survey of long-term agronomic experiments carried out in Italy provided the information in Table 1. They are located at different latitudes from about 37 to 46 °N (Figure 1), and constitute a very interesting source of agro-meteorological data, useful not only for agronomic studies and model building but also for agro-ecological analyses.

Figure 1. Distribution of long term experiments in Italy.

The duration of the experiments varies: the lower limit to the definition of a long-term experiment has been arbitrarily set at twelve years (corresponding to two complete six-year rotations), while the longest running experiment began about 40 years ago. One or more experiments are located on each site. Soft wheat is the most frequent crop in the different locations. Just one location uses a true database for data storage, while data are managed in the others mainly with spreadsheets and hand-written notebooks.

The ClimagriLT data model

To collect, store, and describe data obtained during long-term experiments, a meta-database and a database, integrated in a single application called ClimagriLT, were planned and partly implemented.

The database was chosen because it is a tool offering data protection, easy data sharing and access control (Battini et al., 1986; Atzeni and de Antonelli, 1993; Olson et al., 1999; Van Evert et al., 1999b).

The data model in itself is added information that describes long-term experiments as we perceive them. The organization of the long-term experiments, i.e. locations, treatments, studied factors, data collection methods, availability, distribution, experiment site map are contained in the meta-database. The database stores meteorological data, soil analysis data, crop yields and biometrics. The data model was built following the relational database theory as this is widely accepted, intuitive and easy to implement. Four steps drove the data model construction: i) requirements; ii) draft of the conceptual structure; iii) draft of the logical structure; iv) normalization (Atzeny and de Antonelli, 1993; Atzeni et al., 1996). These steps were reiterated until the data model worked correctly.

- Step I. The "requirements" are a simple list of the objectives guiding the database design and the needs it should satisfy. During the database planning process nothing is blocked in a particular state and requirements are often re-written, using information obtained by the other steps.
- Step II. The phase "draft of the conceptual structure" is the identification of entities that

Table 1. Locations and other information on long-term experiments.

Province	Location	Name	Organization	Begin date
Padova	Legnaro	Legnaro1	University of Padova	1963
Pisa	Pisa	Pisa Sodo Arato	University of Pisa	1989
Matera	Policoro	Policoro	University of Bari	1973
Foggia	Foggia	Foggia	MIPAF (Ministry of Agriculture) - Foggia	1986
Bologna	Cadriano	Cadriano64	University of Bologna	1977
Bologna	Cadriano	Cadriano 29	University of Bologna	1981
Palermo	Palermo	Palermo LT	University of Palermo	1990
Perugia	Papiano	Perugia	University of Perugia	1973
Lodi	Lodi	POC Lodi	MIPAF - Lodi	1989
Foggia	Foggia	Foggia	MIPAF - Bari	1977

usually correspond to "*real-world*" data groups (in long-term experiments a very intuitive entity is the meteo-data group) and the drawing up of the relationships among the entities. Usually, each defined entity turns into a table in the future database (Battini et al., 1986). From the analysis of the entire set of long-term experiments, the plot where each crop was grown represented the key concept of the system. Both from the experimental and the subsequent modelling points of view, the plot, i.e. the physical piece of land, remains the only unchangeable element in experiment history. This approach differs from the agronomic tradition considering the treatment as the most important attribute of an experiment.

Having defined the central role of the experimental plot, the next step was the partitioning in different entities of every group of events that occurs on the single plot. Sowing and harvesting, fertilizations, irrigations, tillage, soil samples become different independent data groups. Linked events like sowing and harvesting are joined in the same entity, while independent events like fertilization, tillage, etc. form separate entities. All the events are directly associated to the plot.

The central role of the plot is also underlined by its relationship with the entity that describes meteorological data passing through the entity that describes experiment locations. Figure 2 summarizes the general model concept.

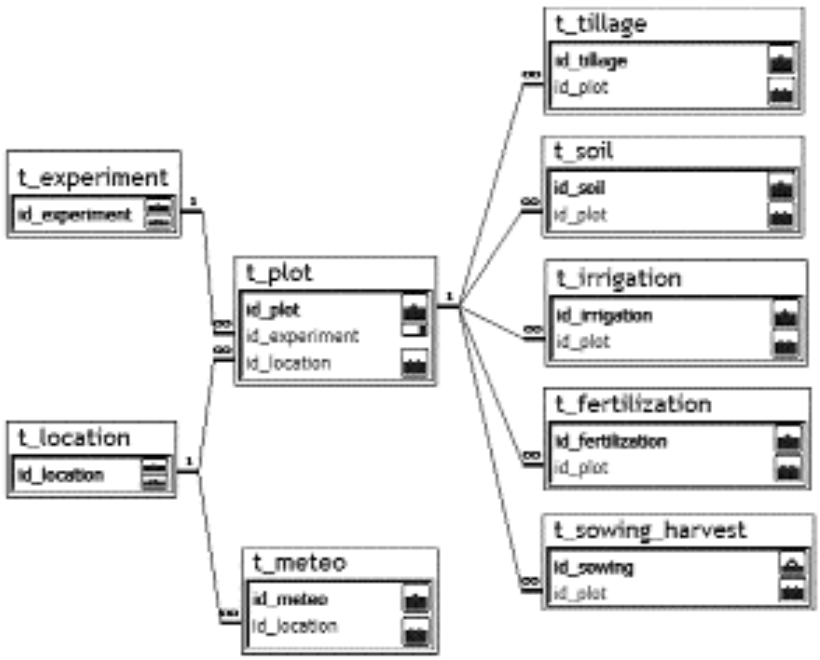

Figure 2. The core data model of ClimagriLT.

Each single plot is also linked to an "experiment" entity, containing metadata on each experiment. The resulting layout can easily be expanded and new entities can easily be added, maintaining plot centrality and improving detail level (an example is given in Figure 3).

During the database feeding tests this data model revealed its flexibility to different data formats encountered in each experiment. In this way the standardization became a direct consequence of database feeding.

- Step III. The phase "draft of the logical structure" is the translation of the conceptual structure in tables and relationships using a Database Management System (DBMS). During the early development of this project we used MSAccess; successively we adopted the freeware MySQL DBMS.
- Step IV. The "normalization" allows to check if the logical structure obtained is coherent, free from any redundancy, with a single updating and deleting point for every data, devoid of non-atomic data. ClimagriLT was normalized at the third normal form. A large database is usually subjected to a conscious and limited de-normalization by planners in order to facilitate its implementation. At the present stage of development ClimagriLT does not require de-normalization, but new requests from data providers, related to easy treatment tracing, will probably require a partial modification of the peripheral structure.

Management policy

ClimagriLT is an application mainly planned to share data from many long-term experiment sources. Data access management is a delicate element to be considered and deserves some organizational effort. A summary of the data management policy is given in Figure 4.

The Database administrator checks data input and is responsible for operation of the software application. In this phase he is not responsible for data quality but only verifies that they come from an authorized source. Read-and-write users (data providers) cooperate with the administrator in feeding data and have free access to the whole metadata set and their own data. Read-only users (i.e., the whole scientific community) have free access to the whole metadata set. A protocol was drafted, signed by the data providers (Table 1), to identify what is metadata and what is not. This protocol contains definitions that were accepted by the organizations managing experiments (data providers). A summary is reported in Table 2.

While data providers have free access to the complete metadata set and to their own data set, the scientific community will have access to single or multiple data sets with the permission of the data providers. The administrator re-directs access inquiries to data owners and, when authorized (by letter, fax, or electronically signed e-mail), gives permission for access.

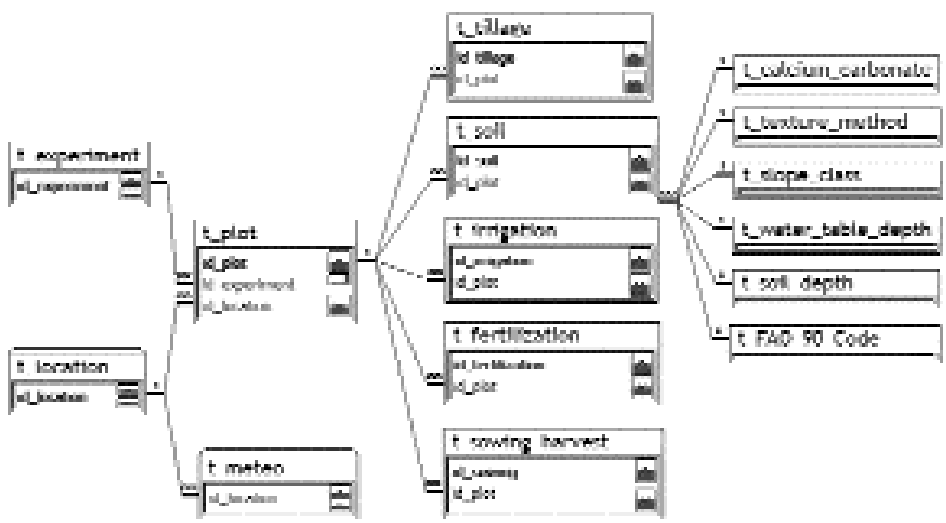

Figure 3. Extension of soil sample entity, adding new entities to improve detail level.

Table 2. Definition, ClimagriLT oriented, of metadata and data.

Metadata	Data
- Organizations managing the experiment - Present head of experiment - Location (lat. long. altitude) - Starting date and end date, if any - Treatments - Crops (species and varieties) - General experimental layout - Plot distribution map - Variables measured - Data format (spreadsheet, dbf,…) - References linked with the experiment - Other data offered - Events occurring on a plot, with info on date and intensity (e.g. Ploughing, 22/10/95, 40 cm; Nitrogen fertilization, 10/03/95, 60 kg/ha) - Meteorological data availability, format and variables measured - Meteorological station (type, history and changes, if any) - Initial conditions availability and completeness of data	- Yield and residue, plus all other crop measured variables during or at the end of growing season. - Meteorological data - Soil analysis data

Present asset of the meta-database

The final infrastructure of ClimagriLT is shown in Figure 5. Its layout derives from our initial ideas and from inputs obtained during creative discussions with data providers and potential users.

ClimagriLT contains two databases: the first collects raw data (BackEnd Database), the other serving simulations (Simulation Server Database, SSD). SSD contains elaborated data obtained by estimates and interpolations, through different tools, in order to fill the gap due to missing data and correct out-of-range data due to input errors (Donatelli, personal communication). All non-original (estimated) data will be marked. The SSD is therefore also a "mirror database" giving further protection against data loss caused by hardware failure. A single computer, working as database server, will be dedicated to each database. A Web server will be activated in a first phase to achieve the minimum task of database publication (Web Server 1). Through this Web Server a user will be able to query the metadata set (and data set if authorized) and communicate with the administrator and data providers. Every access will be registered. A switch addresses the query to the Backend database or SSD. Data providers will have access to the Backend database for data input and checking. Data extracted from the databases will be in XML format in order to be easily converted into other formats and adaptable to recent Internet standards.

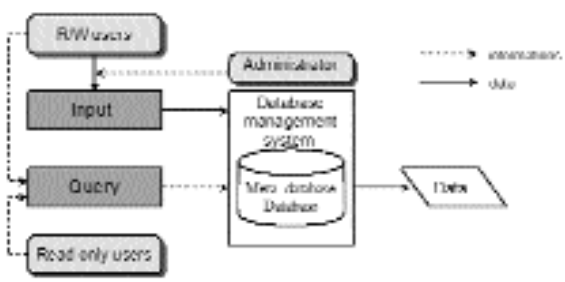

Figure 4. General data management scheme.

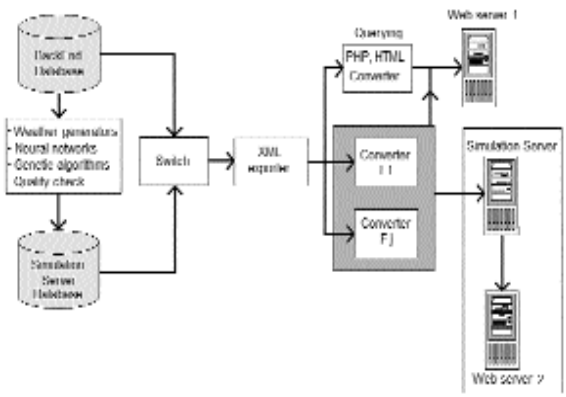

Figure 5. General Infrastructure of ClimagriLT.

Presently, ClimagriLT is implemented at the stage of standalone database, both in Microsoft Access and in MySQL. A set of forms has been implemented for queries and data exporting. The database can easily be explored with Microsoft Access or DBTools DB manager Professional (http://www.dbtools.com.br/). Ten long-

term experiments have already been recorded. Metadata are almost complete and about 20% of the data have been fed into the database (the total number of plots is about 1500). Further work is in progress on problems with the interpretation of older data and initial conditions data, and for data exporting (through a general XML exporter).

Basic aims and future developments

The design, building and feeding of a database like ClimagrLT is the starting point for producing something useful to the scientific community. The two basic aims of ClimagriLT are its publication on the World Wide Web, as a metadata set permitting access for inquiries and cooperation requests, and the creation of an exporter allowing users to extract data ready for different modelling environments.

Database publication on the Internet is a fundamental goal of the ClimagriLT project. The database, particularly in its early development stage, is a dynamic tool, regularly integrated with new data and offering new features and services. Providers of new data could be interested and involved in the project. Inquirers having any kind of access to the meta-database or database will be registered and when using the information in a published paper or other documents they should refer to ClimagriLT and the specific data provider through an appropriate citation, giving proper credit to the scientists involved in long-term experiments. The ultimate goal is to transform the different long-term experiments into a single data source, which can provide a general overview of a large set of agro-meteorological conditions, maintaining specific characteristics and improving the visibility of the single experiments. This approach should lead to improving public use of data and metadata, allowing preparation of value-added services (Doraiswamy et al., 2000), transforming this kind of experiment into a useful public tool.

The capability of data exporting in different formats is also important. We are working on a general exporter, which can create multiple data formats, in order to satisfy a large group of agro-meteorological models. This should facilitate model comparisons and diversify ClimagriLT users. Formats for CSS (Danuso et al., 1999) and Cropsyst (Stöckle et al., 2003) are presently ready and the next step will be the creation of ICASA files (Hunt et al., 2001).

The activation of a simulation server, probably requiring a second web server, will be also considered. This should allow users to manage simulations directly on the net.

ACKNOWLEDGEMENTS

This research is supported by the CLIMAGRI Project of the Italian Ministry for Agriculture and Forestry Policies (MIPAF).

We would like to acknowledge the following scientists for their help and useful advice: Luigi Giardini and Antonio Berti - University of Padova; Marco Mazzoncini and Antonio Coli - University of Pisa; Angelo Caliandro and Vittorio Marzi - University of Bari; Pasquale Martiniello MIPA - Foggia; Giovanni Toderi, Gianni Giordani, Franca Comellini, Loretta Triberti and Anna Nastri - University of Bologna; Luigi Stringi and Luciano Gristina - University of Palermo; Umberto Bonciarelli - University of Perugia; Cesare Tomasoni and Lamberto Borrelli MIPA - Lodi; Vincenzo Rizzo and Michele Rinaldi MIPAF-Bari; Ferruccio Giorgessi and Roberto Carraro MIPA - Conegliano Veneto (TV).

REFERENCES

Acock B., Acock M.C., 1991. Potential for using long-term field research data to develop and validate crop simulators. Agron. J., 83, 56-61.

Atzeni P., De Antonelli V., 1993. Relational database theory. Benjamin Cummings Publishing Company Inc. Redwood City. 389 pp.

Atzeni P., Ceri S., Paraboschi S. e Torlone R., 1996. Basi di dati. Edizioni McGraw-Hill, Milano. 605 pp.

Battini C., De Petra G., Lenzerini M., Cantucci G., 1986. La progettazione concettuale dei dati. Ed. Franco Angeli, Milano, 384 pp.

Codd E.F., 1970. A Relational Model of Data for Large Shared Data Banks Communications of the ACM, Association for Computing Machinery, Inc. 13 (6), 377-387.

Danuso F., Bigot L., Budoi G., Franz D., 1999. CSS: a modular software for cropping system simulation. Proceedings of Agroclimatology and Modelling International Symposium "Modelling Cropping Systems", June 21-23, Lleida, Spain.

Donatelli M., 2003. Personal communication.

Doraiswamy P.C., Pasteris P.A., Jones. K.C., Motha R.P., Nejedlik P., 2000. Techniques for methods of collection,

database management and distribution of agrometeorological data. Agricoltural and Forest Meteorology, 103, 83-97.

Hunt L.A. and K.J.Boote, 1998. Data for model operation, calibration, and evaluation. In: Tsuji G.Y., Hoogenboom G. and P.K. Thornton (eds.): Understanding Options for Agricultural Production, pp. 9-39. Kluwer, Netherlands.

Hunt L.A., 1998. Recent attempts to evaluate and apply wheat simulation models, and to simplify the storage and exchange of experimental data. In: Braun H.J., Altay F., Kronstad W.E., Beniwal S.P.S. and McNab A. (eds.): Wheat: Prospects for global improvement, pp. 445-454. Kluwer, Netherlands.

Hunt L.S., White J.W., Hoogenboom G., 2001. Agronomic data: Advances in documentation and protocols for exchange and use. Agricoltural Systems, 70, 477-492.

Jones J.W., Hoogenboom G., Porter C.H., Boote K.J., Batchelor W.D., Hunt L.A., Wilkens P.W., Singh U., Gijsman A.J., Ritchie J.T., 2003, The DSSAT cropping system model, European Journal of Agronomy, 18, 235-265.

Stöckle C.O., Donatelli M., Nelson R., 2003. CropSyst, a cropping system simulation model. European Journal of Agronomy, 18, 289-307.

Olson R.J., Briggs J.M., Porte J.H., Mah, G.R., Stafford S.G., 1999. Managing data from multiple disciplines,scales, and sites to support synthesis and modelling. Remote Sens. Environ. 70, 99-107.

Van Evert F.K., Spaans E.J.A., Krieger S.D., Carlis J.V., Baker J.M., 1999a. A database for agronomical research data I: data model. Agron. J. 91, 54-62.

Van Evert F.K., Spaans E.J.A., Krieger S.D., Carlis J.V., Baker J.M., 1999b. A database for agronomical research data II. A relational implementation. Agron. J. 91, 62-71.

CLIMAGRILT: UN META-DATABASE RELAZIONALE PER LA GESTIONE DEI DATI DI ESPERIMENTI AGRONOMICI DI LUNGA DURATA

SCOPO. La progettazione, la calibrazione e la validazione dei modelli matematici dei sistemi colturali dovrebbero basarsi su numerosi dati ottenuti da esperimenti agronomici di lunga durata. L'utilità degli esperimenti condotti in Italia è fortemente limitata dalla scarsità di informazioni facilmente disponibili sulla loro organizzazione e sulle modalità di raccolta dei dati e, inoltre, da difficoltà nell'accesso ai dati stessi.
METODO. In questo lavoro viene presentata la filosofia e il disegno generale di un meta-database (ClimagriLT), progettato per immagazzinare, e distribuire i dati relativi a un insieme di esperimenti di lunga durata (trattamenti, produzioni, terreni, clima etc.). Particolare rilevanza è stata attribuita al modello dei dati e alla politica di diffusione dei dati. Il meta-database contiene l'organizzazione degli esperimenti (localizzazione, trattamenti, fattori allo studio, metodi di raccolta dei dati, disponibilità, mappe degli esperimenti; il database raccoglie dati meteorologici, analisi dei terreni, rese e dati biometrici delle colture. Il modello è stato costruito seguendo la teoria relazionale, ampiamente accettata, intuitiva e di facile implementazione. La costruzione del modello comprende i seguenti stadi: i)definizione degli obiettivi; ii) disegno della struttura concettuale; iii) disegno della struttura logica; iv) normalizzazione. Vengono delineati, infine, possibili sviluppi futuri.

Parole chiave: modello dei dati, dati condivisi, definizione di metatdato, serie temporali.