Ricerca svolta nell'ambito del progetto finalizzato CLIMAGRI
finanziato dal Ministero delle Politiche Agricole e Forestali
D.M. 639/7303/2003 - Pubblicazione n. 19

ClimagriLT: a relational meta-database for data management of long term agronomic experiments

Running title: Data management of 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

In press: Italian Journal of Agronomy M.Zuliani, A.Peressotti, G.Zerbi, G.Zuliani, G. Delle Vedove and F.Danuso (2003) ClimagriLT: a relational meta-database for data management of long term agronomic experiments. It. J. Agronomy, 7,2 (106-111)

Abstract

Data obtained from long term agronomic experiments are a basic starting point for development, calibration and validation of mathematical models of cropping systems The usefulness of these experiments carried out in Italy is strongly impaired by scarce information on easily available data aboput their organization and data collecting methods and by difficulties in their access.

The philosophy and the 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 . Focus is particularly set on data model and management policy of data distribution. The organization of the long term experiments, i.e. locations, treatments, factors studied, data collection methods, availability, distribution, experiment site map are contained in the meta-database. The database stores meteorological data, soil analytical data, yield and biometrics of crops. Data model was built following the relational database theory because it is widely accepted and because it is 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 to future applications is also given.

ClimagriLT: un meta-database relazionale per la gestione dei dati di esperimenti agronomici di lunga durata

Riassunto

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 scarse informazioni facilmente disponibili sulla loro organizzazione e sulle modalità di raccolta dei dati e, inoltre, da difficoltà nell'accesso ai dati stessi.

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, trattementi, 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.

Introduction

Studies on agroecosystems are a long term challenge in the same way that agriculture is a long term enterprise. Shifting time perspective from short to long term represents is essential to understand the dynamics of these manged ecosystems that, as a consequence of land exploitation and human pressure rise, cover a relevant fraction of the planet (Army et al ,1991).

Traditional agronomic experiments and empirical knowledge of agreoecosystems are not sufficient to fill the complex matrix of indicators that decision makers require to manage the whole agricultural system at regional and farm level (Jones et al, 2003). It is widely accepted that a possible approach to shift time scale from short to long term and improve knowledge on agroecosystems is represented by mathematical models. Data obtained from long term agronomic experiments are a basic starting point for development, calibration and validation of reliable models aimed at improving their prediction capability and at detecting lacks of knowledge and research activity (Acock et al ,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 represented by lack of a common standard in data storing and managing (Hunt 1998 and Hunt et al. 1998). Long term experiments frequently offer a series of problems deriving from inconsistecies due to changes in the layout of the experimental designs, to the transfer of the responsibility from a scientist to another, to reductions of the required funds etc. Collecting the complete documentation of a single long term experiment is often difficult: loss of data, changes in analytical 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 often probable.

A number of scientists involved in projects on different research fields (plant breeding, modelling, agrometeorology, soil science, plant phisiology) begun quite recently to work on this topic to find solutions to a more efficient data storing and managing (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 were also tested. This last method was used by IBSNAT (International Benchmark Sites Network for Agrotechnology Transfer) developing a set of text files that were recently upgraded by the International Consortium for Agricultural System Application (Hunt et al 2001).

The philosophy and the 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 . Focus is set on data model and management policy of data distribution. A brief introduction to future applications is also given.

Agronomic long term experiments in Italy

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

Tab 1. Location and other information of 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	1985
Foggia	Foggia	Foggia	MIPAF - Bari	1977

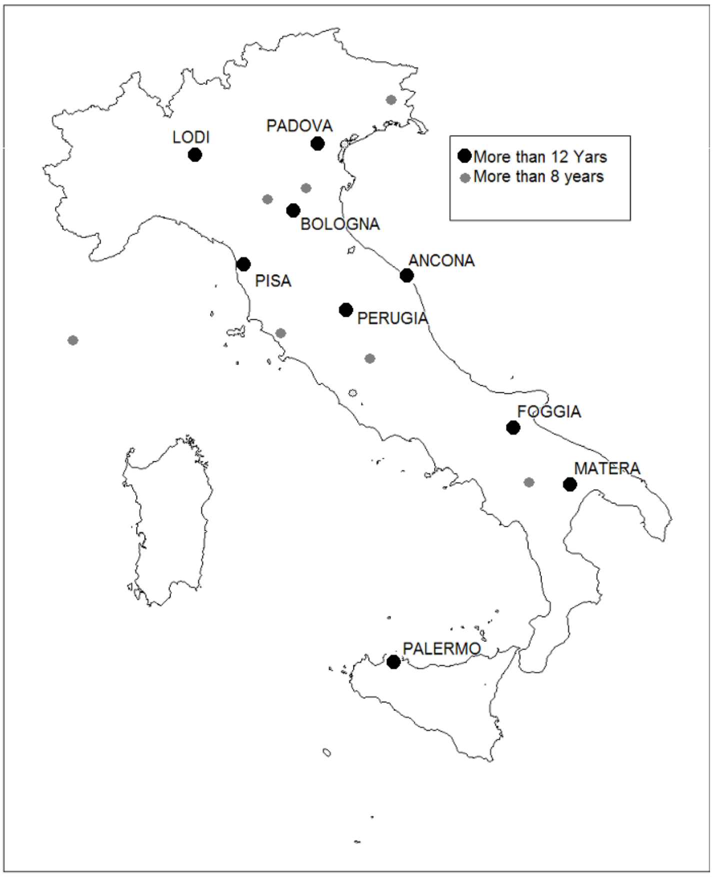

Fig 1. Distribution of long term experiments in Italy

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

The data model of ClimagriLT

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 partially implemented.

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

Data model by itself is an added information that describe long term experiments as we perceive them. The organization of the long term experiments, i.e. locations, treatments, factors studied, data collection methods, availability, distribution, experiment site map are contained in the meta-database. The database stores meteorological data, soil analytical data, yield and biometrics of crops. Data model was built following the relational database theory because it is widely accepted and because it is 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 (Atzeni et al. 1993, 1996). These steps were reiterated until the obtained data model worked correctly.

Step I. Requirements are a simple list of the objectives guiding the database design and the needs that it should satisfy. During 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 usually correspond to "*real-world*" data groups (in long term experiments a very intuitive entity is represented by meteo-data group) and in the drawing up of the relationships among the entities. Usually, each defined entity turn into a table in the future database (Battini et al. 1986). From the analysis of the whole set of long term experiments, the plot where each crop treatment was grown represented the key concept of the system. Both from the experimental and from the following modelling points of view, the plot, i.e. the physical portion 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.

Defined the central role of the experimental plot, the following step was the partition in different entities of every group of events that occurs on the single plot. Sowing and harvest, fertilizations, irrigations, tillage, soil samples turn into 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 linked with the plot.

The centrality of the plot is also underlined by its relationship with the entity that describes meteorological data passing through the entity that describe experiment locations. Fig. 2 summarize the general model concept.

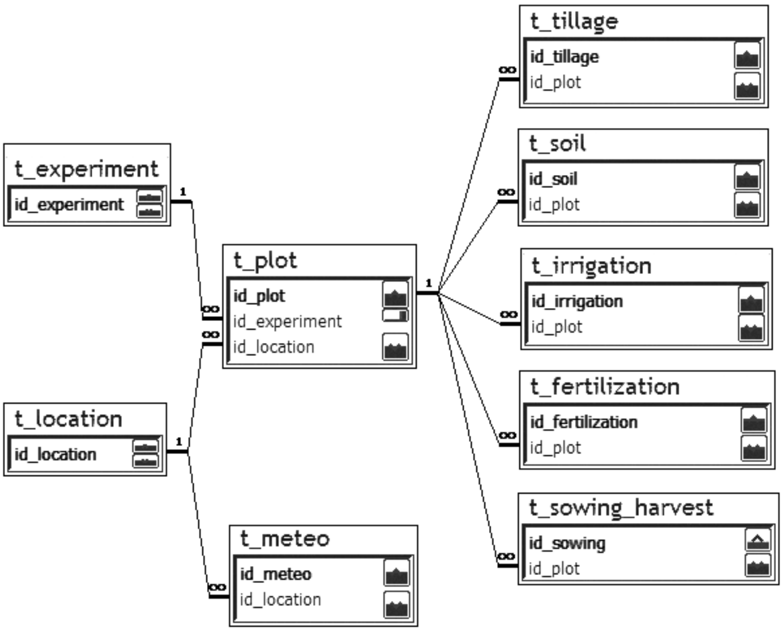

Fig.2 The core data model of ClimagriLT

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

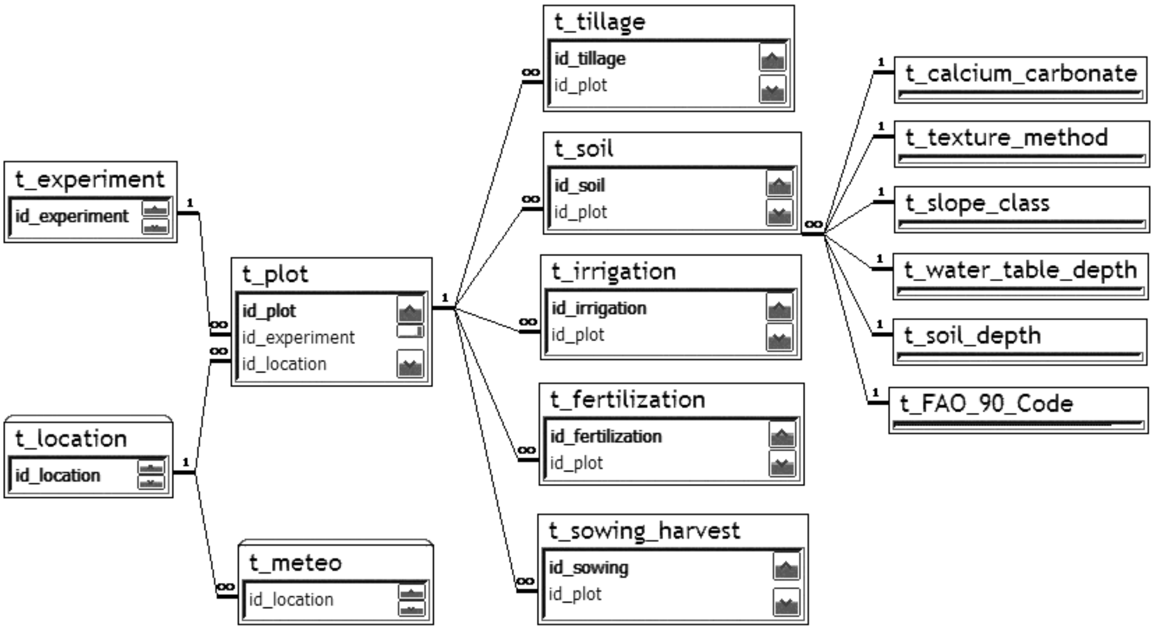

Fig.2 Extension of soil sample entity, adding new entities to improve detail level.

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

Step III. This 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. 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. Usually a big database is 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 need de-normalization but new requests asked by data providers, related to an easy treatment tracing, will probably require a partial modification of the peripheral structure.

Management policy

ClimagriLT is an application mainly planned to share data coming from many long-term experiments sources. Data access management is a delicate element to be considered and deserves some organizing effort. A summary scheme of data management policy is shown in Fig. 4 .

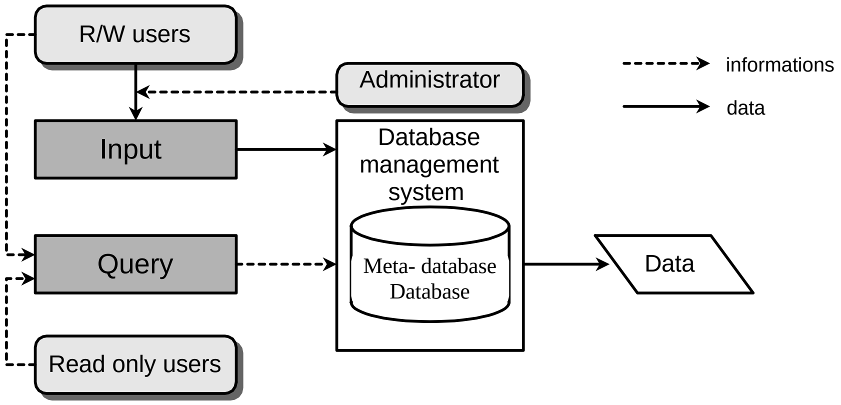

Fig 4. General data management scheme

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

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

Metadata	Data
• Organizations that hold experiments • Experiment responsible • Location (lat. long. altitude) • Begin date and end date, if any. • Treatments (how many and typology) • Crops (species and varieties) • General experimental scheme • Plot distribution map • Variables measured • Data format (spreadsheet, dbf,…) • Reference linked with experiment • Other data distribution • List of event occurring on a plot, completed by date and intensity (eg. 22/10/95 Plowing 40 cm. eg. 10/03/95 Nitrogen fertilization 60 kg/ha) • Meteorological data availability, format and variables measured • Meteorological station (type, story and changes, if any) • Initial conditions availability and integrity data	• Yield and residue, together with all other crop measured variables during or at the end of growing season. • Meteorological data • Soil analysis data

While data providers have free access to the complete metadata set and to their own data set, the scientific community (read-only users) will have access to single or multiple data sets after 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 to the access.

Present asset of the meta-database

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

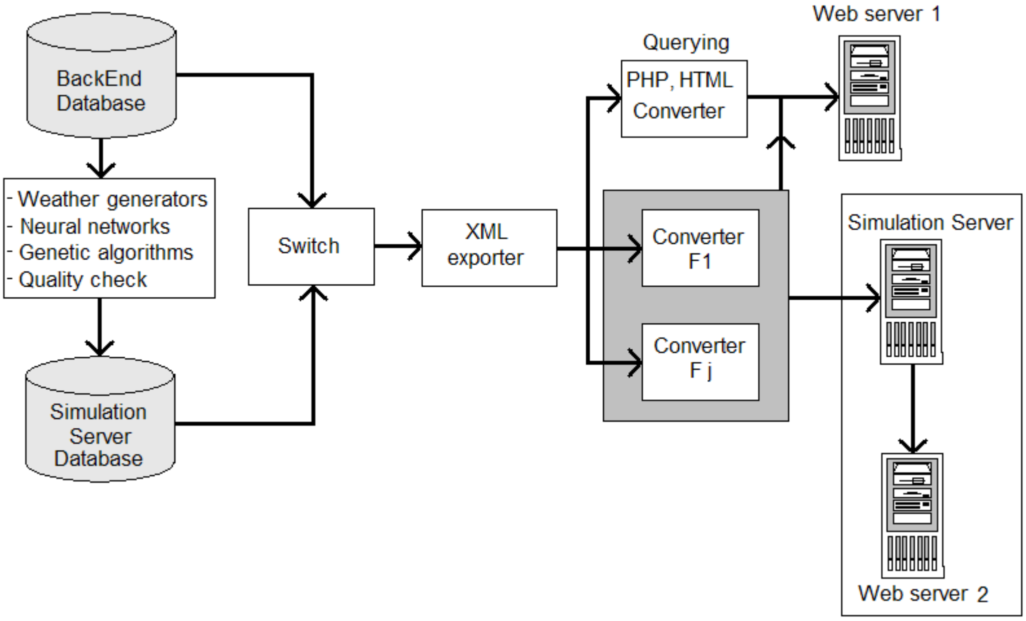

Fig.1 General Infrastructure of ClimagriLT.

ClimagriLT contains two databases: the first collects raw data (BackEnd Database), the other serving simulations (Simulation Server Database, SSD). SSD contains elaborated data obtained with estimates and interpolations, through different tools, in order to fill the gap due to data missing and correct out of range data due to input errors (Donatelli, 2002, Personal Communication). Every non-original (estimated) data will be marked. Therefore the SSD is also a “mirror database” further protecting from 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 metadata set (and data set if authorized) and communicate with the administrator and data providers. Every access will be registered. A switch address the query to Backend database or SSD. Data providers will have access to the Backend database for data input and data check. Data extract from databases, will be in XML format in order to be easily converted in other formats and adaptable to recent Internet standards.

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 for data exporting. The database can be easily explored with Microsoft Access or DBTools DB manager Professional (http://www.dbtools.com.br/). Ten long term experiments have been already recorded. Metadata are almost completed and about 20% of the data have been feeded into the database (the total number of plots is about 1500). Further work is in progress about 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 designing, building and feeding of a database like ClimagrLT is the starting point that allows to produce 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 related to 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 project ClimagriLT. The database, particularly in its early development stage, is a dynamic tool, often 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 the database will be registered and when using the information in a published paper or other documents they should refer to ClimagriLT and to the specific data provider through an appropriate citation, giving the right credit to the scientists involved in long term experiments.

The ultimate goal is to transform the different long term experiments in a single data source, capable to give a general overview on a large set of agrometeorological conditions, maintaining specific characteristics and improving the visibility of the single experiments. This approach should lead to improve the public use of data and metadata, allowing preparation of value-added services (Doraiswamy et al., 2000), transforming this kind of experiments in a tool of public usefulness.

The capability of data exporting in different formats is also important; we are working on a general exporter, able to create multiple data formats, in order to satisfy a large group of agro-meteorological models. This should facilitate model comparisons and extend 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

We would like to acknowledge the following scientists for help and useful advices:
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 MIPAF, 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 MIPAF - Lodi; Vincenzo Rizzo and Michele Rinaldi MIPAF- Bari

References

Acock B., Acock M.C., 1991. Potential for using long-term field research data to develop and validate crop simulators. Agronomy Journal, 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 ed. 1999. 605 pp.

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

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 (2000)

Hunt L.A. and K.J.Boote, 1998. Data for model operation, calibration, and evaluation. In 'Understanding Options for Agricultural Production', eds. Gordon Y.Tsuji, Gerrit Hoogenboom and Philip K.Thornton. 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 'Wheat: Prospects for global improvement', eds. H.J. Braun, F. Altay,W.E. Kronstad, S.P.S. Beniwal and A. McNab. 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 (2001): 477-492

Stöckle C.O., Donatelli M., Nelson R., 2003. CropSyst, a cropping system simulation model. European journal of Agronomy, 18 (2003): 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 (1999)

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.