Integration of Heterogeneous Pluviometric Data For Crop Forecasts

JOÃO GUILHERME DE SOUZA LIMA[1]
CLAUDIA BAUZER MEDEIROS[2]
EDUARDO DELGADO ASSAD[3]

[1,2]Instituto de Computação, Universidade de Campinas - Caixa Postal 6176, 13081-970, Campinas, SP, Brasil
{joao.guilherme, cmbm}@ic.unicamp.br
[1,3]Embrapa Informática Agropecuária - Av. Dr. André Tosello, nº 209, 13083-886, Campinas, SP, Brasil
assad@cnptia.embrapa.br

Abstract. Crop forecast is an activity practiced by experts in agriculture, based on large data volumes. These data cover climatological information of the most diverse types, concerning a geographic region and the type of culture. Besides volume, another problem to face concerns data heterogeneity. This paper presents a project for development of a data management system for crop forecasts. The paper is centered in the management of pluviometric data, an important factor in crop management. The system is being implanted by Embrapa, the Brazilian Agricultural Research Corporation, and part of it is already available on the Web.

1 Introduction

Due to its random character, pluviometric precipitation is one of the factors that influence agricultural production the most, increasing the risks of scheduling activities [5]. Today, damages originated from crop losses due to rains, frosts or droughts vary between R$500,000 and R$9,000,000 annually in Brazil. In specific cases, such as the corn crop, losses have already reached R$120,000,000 per annum.

Crop losses can be reduced by collecting and analysing climatological data, which allow farmers and cooperatives to better schedule their activities. The Brazilian Agricultural Research Corporation, Embrapa, is developing a Web-based agroclimatological monitoring system for the entire country. This system will make available on the Web several kinds of information about rain, water availability in soil, mean temperature and soil handling conditions. This information – the system's *products* – will be provided under several formats, e.g., maps and tables. End users (cooperatives, agronomists) will be able to utilize such products as basis for decision taking in different phases of a crop management cycle, assisting, for example, the scheduling of planting, harvesting and drying; the application of agricultural defensives and the preventive control ou direct combat to frosts.

One kind of data to be handled by the system are historical series of pluviometric measures. Such series allow estimation of the probability of rain and sequential occurrences of rainy and dry periods in specific regions. By studying the rain patterns of a region, one can determine locations with similar pluviometric conditions and conveniently group them in sub-regions. This classification, combined with a characterization of the kind of soil, allows to determine which kind of culture is more adequate, when is more convenient to plant it and what are the associated risks. There are several institutions in Brazil that gather pluviometric data. However they do not share their data among themselves.

Better forecasting requires gathering and integrating these data. The integration of pluviometric data in Brazil presents several problems originated by their heterogeneity. Many factors contribute to this scenario. First of all, data are distributed along diverse institutions. Each source responsible for effectuating pluviometric measures has defined its own ways of storing data. The Agriculture Ministry, Agricultural Institute of Campinas, Water and Energy Department of state of São Paulo and the Water National Agency are examples of such sources. While some of them manage their data using DBMS, others keep data in Excel sheets or even in text files, managed by legacy systems. Moreover, data are collected by different devices, with distinct periodicities and precisions. Usually there are gaps in the historical series – e.g., due to device or reading errors. Therefore, to reach an efficient integration it is necessary to deal with questions such as data homogeneization, data consistency and historical series processing and management, where the geographical location plays an important role.

This paper describes the efforts towards integrating these data within a project being conducted by Embrapa. The main contributions of the paper are the description of integration problems for these kinds of georeferenced data and the analysis of the adopted solution, which combines temporal series treatment, statistical and geostatistical processing and heterogeneous data management.

The rest of this text is organized as follows: section 2

reviews concepts necessary to the work, analysing aspects of heterogeneous data integration, data quality evaluation and pluviometric data. Section 3 presents the proposed approach and its implementation, and section 4 concludes the work.

2 Review of Related Literature

The solution presented requires combining work in spatio-temporal data, heterogeneity issues, quality evaluation and pluviometric data. This section presents basic concepts related to the work, and discusses some of the solutions proposed for dealing with these problems.

2.1 Spatio-temporal databases

The data treated in this work are *spatio-temporal*: they have *spatial* attributes (which define their geographical features), *temporal* attributes and *descriptive* attributes. The latter can vary even if spatial attributes do not, and vice versa.

In the context of this paper, the main kind of spatio-temporal data to be handled consists of daily pluviometric measures. Such measures are collected by pluviometric *stations*. The spatial attribute is the station's geographic location (latitude, longitude). A descriptive attribute is the measure of the amount of rain in a specific date, and the temporal attribute consists of the date itself. Section 2.4 qualifies pluviometric data more precisely.

The main problems in this context are the irregular sampling (either spatial or temporal) and the existence of time periods without measures recorded (measurement gaps). The problem of obtaining these missing attribute values comes into play (the so called *temporal interpolation*). *Spatial interpolation* on the other hand tries to supply geographic data for regions without such data. In both cases, one can utilize functions that try to deduce the missing value taking as basis other periods or locations for which there are records in the spatio-temporal database. The development of such functions is not trivial and can present many difficulties [22]. The data considered by this work are subject to both spatial and temporal gaps, and demand solutions with spatio-temporal interpolation. In geographic applications, frequently an attribute value is considered to be constant from the time for which it is valid until the next timestamp stored. This approach is not adequate in most cases for pluviometric measures, and thus other kinds of interpolation had to be applied, as will be seen.

2.2 Some heterogeneity issues

Heterogeneous database environments have been defined in the literature in various ways and levels. The differences can occur among DBMS, among data models, among schemas or in the data semantics. These differences can also occur in the hardware (platforms).

The resolution of semantic conflicts resides in defining and standardizing the meaning of concepts, terms and structures found on data sources. As evidenced in [30], this is the most problematic heterogeneity level. A current research field is the *semantic Web* development [6, 12], which aims to increase the information sharing on the Web making data semantics more machine-understandable. In such field, *ontologies* [11, 24] have been proposed as a way to organize knowledge in a structured way to allow interoperability [13].

In the spatio-temporal data context, one can solve conflicts by integrating models, schemas or data. The final level of data integration is usually based on the coordinates (the spatial component), but this can also cause semantic problems. Another factor to consider is the *temporal granularity* – the time unit chosen as basis for recording changes. This choice depends of the intended applications. Specific values can be measured or supplied with different periodicities (for example, day, month, year or season).

The data considered in this work present heterogeneity problems in platform, model and schema. They also appear with distinct temporal granularities. As will be seen, the solution proposed involves using coordinates for spatial integration and interpolation, mapping the data into a single schema, and imposing rules on temporal conversion. A *XML Schema* [9] and *metadata* [18] are specified to foster interoperability [4, 25, 14].

2.3 Data Quality

The question of whether to use a given database is associated with a judgment that can best be defined as quality, or *fitness for use*: "does this information source serve my purposes sufficiently to be worth the effort to obtain it and convert it into my analysis?" [7]. This question is especially appropriate in the context of this work, given the spatio-temporal heterogeneity characteristics of the treated data. Guidelines to quality assessment in geographic systems can be found in [23].

Hohl [15] identifies two distinct quality processes in geographic data conversion: quality control and quality assurance. *Quality control* verifies the way data are collected and managed, trying to ensure that they always reach a quality level initially established. *Quality assurance* consists of the final verification of the converted data, before being stored on the database. Besides verifying the final quality level of data, this process also identifies and corrects errors. The described system deals with quality assurance, while, ideally, quality control must be made by the data sources.

Geographical data storage requires sampling, which can introduce errors. Thus, one must establish acceptance

criteria for data sets. Another problem happens when more than one source supplies the same data, since in this case one must choose the appropriate source. It is thus necessary to quantify quality, using for instance distinct quality *indicators*.

Indicators can be used to analyse quality under a multi-dimensional point of view. Here *dimensions* are specific properties of data or data sets for quality purposes. There are distinct proposals for defining reference sets for such dimensions. Mecella et al. [21] propose four dimensions: accuracy, completeness, timeliness and inner consistency. Pipino et al. [27] enumerate 16 different dimensions. The choice of dimensions depends on an application context.

Whether or not many indicators are used, quality evaluation may be task dependent or not. *Task-independent* evaluation metrics reflect data characteristics without contextual knowledge of the application, and can be applied to any data set, regardless of the tasks to be executed. *Task-dependent* metrics must be developed in specific application contexts [27]. They can include business rules, company regulations and constraints determined by the database administrator.

In the case of the work related here, the large amount of data sources and the existing gaps in temporal series demanded association of quality dimensions to the integrated data. Critical decisions are taken based on the informations returned by the system, increasing the necessity for quality parameters.

2.4 Pluviometric Data

The amount of rain that occurred in a place is recorded by means of two kinds of devices: the pluviometer and the pluviograph. Measures are periodically taken, usually in a daily basis.

Pluviographs are mechanical devices that produce precipitation graphs, generating errors only if there is some mechanical deviation in the device. In the case of pluviometers, data readings are made by human operators, who can make mistakes. Reading and transcription errors and also omissions of the operator and time periods in which the station do not operate can produce gaps or wrong records in a historical series.

This requires applying consistency methods to historical series before their data are utilized. Such methods try to detect (and possibly correct) unreliable values and to fill gaps. These actions aim to ensure that the series attain a minimum quality level.

Some examples of consistency methods for historical series of pluviometric measures are the *double-mass curve* technique, *Principal Component Analysis* [19], *Time Interpolation of the Principal Component Scores Series, Penalty of the Principal Component Scores* [20], *Regional Vector Method* [29] and *Nearest Neighbor Method* [20]. Such methods are based on geostatistical processings. They commonly assess and correct data provided by a weather station using data from neighbor stations.

The pluviometric data supplied to the system discussed in this paper can be stored by the source institutions in diverse formats and kinds of files. Figures 1 and 2 exhibit two examples of such data files – Figure 1 shows data provided in Access format, while Figure 2 shows an ASCII format. The figures show that not just the kind of storage differs, but also the recorded attributes. As will be seen next, this required establishing a conversion standard to integrate those data.

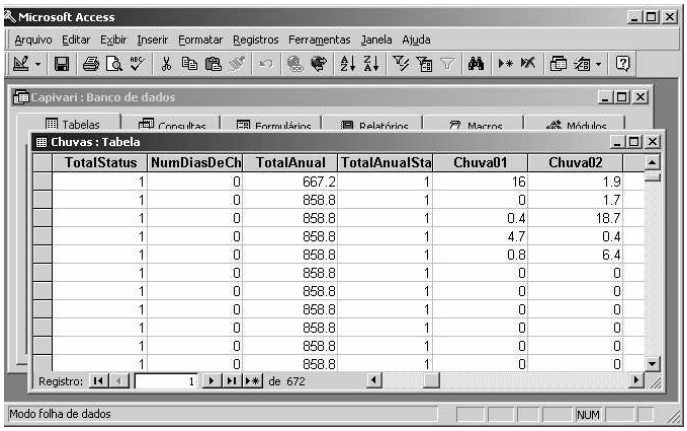


Figure 1: *Data file example - Access format*

```
AMERICANA 224200 471700 540.0
D4-004,1978
3 0 52 0 0 3 136 0 114 5 2 0 352 0 0 0
0 78 0 0 0 0 120 0 0 0 0 0 0 0 0 4
0 0 0 0 60 0 525 0 0 0 0 142 5 8 0 0
0 12 0 0 60 0 193 102 0 0 0 0 207 0 0 0
223 0 20 113 72 0 0 585 62 55 38 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 10 25 0
25 79 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 264 54 36 0 116 0 0 0 0 0 0 0 0 12 0
3 0 0 0 0 0 0 0 655 6 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
385 0 48 8 0 122 69 379 0 15 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 4 0 6 0 0 0 0 0 0 0
```

Figure 2: *Data file example - ASCII format*

3 Architeture and system overview

The solution proposed is based on the assumption that all data will be merged in a centralised system. Heterogeneous data received from the various sources are initially standardized, integrated to a database with associated metadata, and then submitted to a quality evaluation process. Data,

metadata and quality indicators are managed in a Oracle relational DBMS.

To reach the objectives of integration, quality evaluation and product supplying, the system was divided into the modules shown in Figure 3: A) Data Integration, B) Quality Evaluation and C) Querying Processing modules.

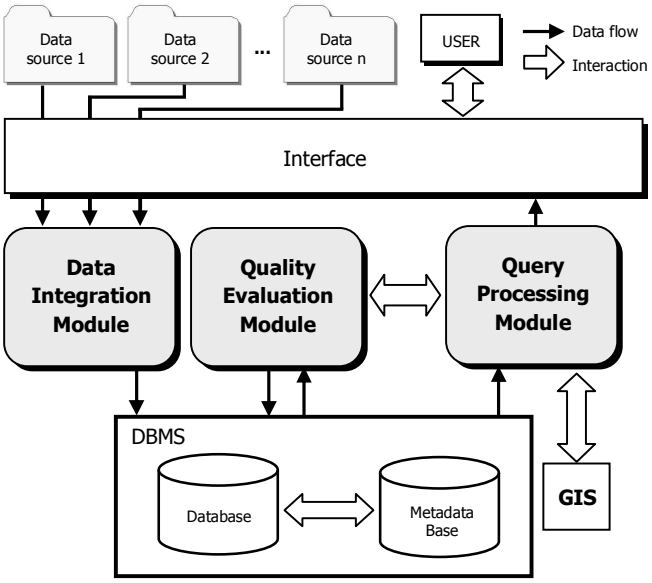


Figure 3: *System architeture.*

The database stores pluviometric series, maps and additional data about the desired regions (for example, kind of soil). The relational schema established to integrate pluviometric series allows subsequent data treatments that one could require to apply. The metadata describe the data on the database, as explained next.

Metadata base

The goal of the metadata base is to help the user in determining which data sets serve his purposes. These metadata will also help interoperability, when the system´s integrated data are supplied to other systems. Having this in mind, the metadata describe characteristics that commonly cause heterogeneity problems between climatological data sets.

The set of metadata attributes was designed considering the following standards: *Content Standards for Digital Geospatial Metadata* (CSDGM) [10], from Federal Geographic Data Committee (FGDC); *WMO Core Metadata Standard* [26], from World Meteorological Organization (WMO) [2]; *Metadata to Scientific Workflows to Support Environmental Planning* [28] and *Dublin Core Metadata Initiative* (DC) [1]. The first three were chosen since they describe data whose scope is related with climatological data scope. The DC stantard was used in naming metadata elements. This leads to interoperability, since the semantics of each DC element is becoming a world standard.

The resulting metadata attribute set comprises data Identification, Spatial Coverage, Temporal Coverage, Units and Management Issues. Therefore, the metadata base includes, among others, information about the data sources, pluviometric stations, stored maps and processes the data were previously submitted to. The complete metadata set can be obtained at [17].

Besides the cited metadata standards, the solution adopted metadata *codification* conventions proposed in the literature. The adoption of such standards and controlled vocabularies decreases heterogeneity and eases data acceptance by other systems.

A) Data Integration Module

The crop forecast system concerns integration of various data types (climatological, geographical and cultures). This paper focuses on the integration of pluviometric data. Input pluviometric data are provided by institutions all over Brazil. Some are provided once a year (historical series), while others are daily supplied with up-to-date data from several stations. Moreover, legacy series, some with 100 years of data, are also being imported into the system. The Data Integration Module is responsible for formating these data into a unique standard.

Each data provider has its own way of collecting and storing data, originating problems like divergences on precision, frequency of sampling, and quality of data and on the supplied metadata. Standard techniques for heterogeneous data integration suggest two alternatives: real integration, converting data to a standard model and then integrating them; and virtual integration, in which data are not modified, but a set of mediators offer an integrated vision. The first solution was favored, since the goal is to keep processed data in a data warehouse. Data producers can keep their data in the original formats and schemas.

Each source's data conversion are made by *data wrappers*. Once converted, data are submitted to a *Data Migrator*, which inserts them in the database. Figure 4 exhibits the Data Integration Module in a schematic way.

Wrappers convert the data to a single format, resolving heterogeneity problems related to temporal granularity, units, precision and inferred data. A wrapper for each data source was implemented, for storage formats like ASCII, Excel, Access and DBF.

XML was the chosen output format for wrappers. This makes exchange easier if data are provided to others systems. Although GML [8] facilitates geographic data acceptance, it was necessary to define a specific XML Schema [9] to validate the XML documents, due to various particular data and metadata attributes inherent to the climatological

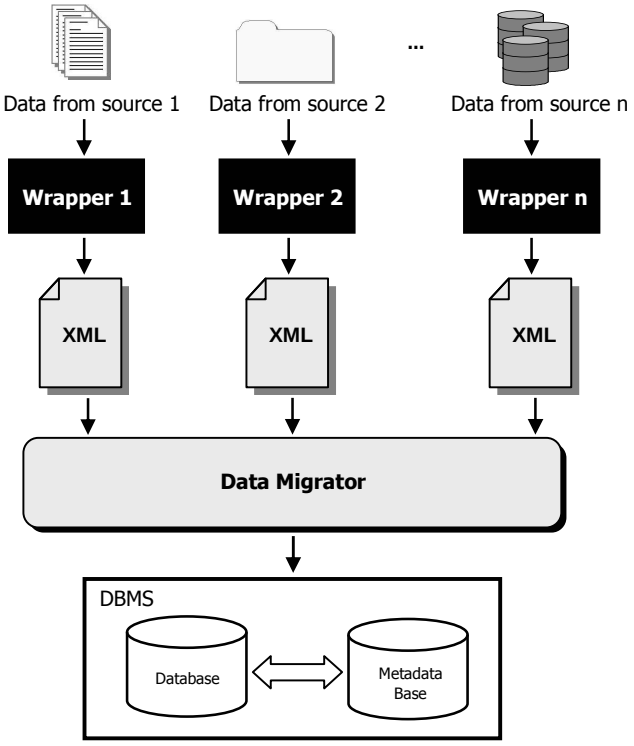


Figure 4: *Data Integration Module.*

data in hand. This schema is available in [16].

The basic function of the Data Migrator is to read the XML documents generated by wrappers and to insert the converted data in the integrated database. The migrator also checks some value space constraints, updates data for consistency, verifies whether a measure already exists in the integrated database, and deals with the problem of identifying, in the database, the station of each measure supplied.

B) Quality Evaluation Module

Most of the queries on climatological data require geostatistic processing to generate consolidated information – e.g. spatially referenced data to be visualized in maps. Results of quality evaluation help the user in identifying adequate data sets, and supply the user with parameters of the result's credibility. The Quality Evaluation Module performs such assessment. All consistency and scoring criteria applied in the analysis were established after discussions and research with agroclimatological experts.

Quality evaluation is divided in two phases. The first one is a consistency phase, where data are analysed and, when possible, corrected to become more adequate to query processing. The second phase associates quality indicator scores to specific time periods. The consistency phase has three stages.

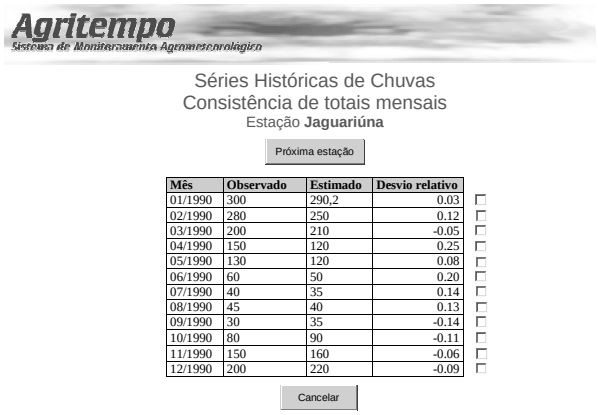
The first stage analyses a historical series, verifying its extension and existence of gaps. According to pre-defined criteria, some years of the series or even the entire series can be "rejected" in the analysis. Rejected data are not deleted from the database, they are just not made available for queries. This stage also processes daily data, which can be marked as *unreliable*. A given measure is considered as such if it is greater than 150mm or if it is greater than the critical value of the corresponding homogeneous region and day of the year. In this case, *homogeneous regions* are geographical regions defined according to pluviometric behaviour. Each region has a 3% critical value, for each day of the year. The possibility of occurrence of an amount of rain greater than that indicated by the critical value, in the given region and day of the year, is less than 3%. Critical values and homogeneous regions composition are not calculated by the system, but defined by the Statistical Group of the Agricultural Monitoring Project of the Agriculture Ministry.

After this analysis, the next stage is inference of missing daily data. The inference method uses data from neighbor stations, applying a weighed average based on the inverse of the square of the distance. Measures calculated in this phase are marked as *inferred*. Figure 5 shows the screen where the data administrator can ask for inference of missing measures.

Figure 5: *Inferring missing data - screen copy.*

In the last stage of the consistency phase, pluviometric totals of months are analysed through the *regional vector* statistical method [29]. This method identifies a pluviometric behaviour for a region and period of time, and provides data administrators – agricultural experts – with the possi-

bility of storing new inferred totals. Figure 6 exhibits the screen where the administrator compares the observed and inferred totals for a specific station, and specifies which inferred totals must be stored. Notice that when estimated totals are recorded, the corresponding original data are also kept in the database.



Mês	Observado	Estimado	Desvio relativo	
01/1990	300	290,2	0.03	☐
02/1990	280	250	0.12	☐
03/1990	200	210	-0.05	☐
04/1990	150	120	0.25	☐
05/1990	130	120	0.08	☐
06/1990	60	50	0.20	☐
07/1990	40	35	0.14	☐
08/1990	45	40	0.13	☐
09/1990	30	35	-0.14	☐
10/1990	80	90	-0.11	☐
11/1990	150	160	-0.06	☐
12/1990	200	220	-0.09	☐

Figure 6: *Screen for consistency of month pluviometric totals.*

The second phase of the quality assessment, where data are associated with scores, performs two kinds of evaluation: *task-independent* and *task-dependent*. All indicators used in this phase are quantitative.

The task-independent evaluation analyzes data independently from the intended use. It is performed after the consistency phase, and its results are stored and used in all queries. The quality indicators applied in the task-independent evaluation are *completeness* and *free-of-error*. Each **year** of series previously approved in the consistency phase receives one score for these indicators, normalized in a 0-1 scale. The score for completeness is based on the amount of missing daily data in the year, using a direct mapping: a year with 15% missing data receives a completeness score of 0.85. This direct metric allows the user to make subsequent specific analyses, which would not be possible if a more sophisticated method had been used. The score for free-of-error is determined through the amount of ocurrence of *unreliable* daily data in the year, previously calculated in the consistency phase.

Task-independent evaluation is executed automatically after the consistency phase, and it is re-executed periodically to reflect changes in the database. Results of the task-independent evaluation are stored in relational format in the DBMS, and are used in queries, jointly with metadata.

Task-dependent evaluation is driven by queries. It utilizes the *timeliness* and *appropriate amount of data* quality indicators. The first one represents the extent to which data are sufficiently up-to-date for the task being performed, and the second represents the extent to which the volume of data is appropriate for the task in hand. Since they are task-dependent, the metrics for these indicators are determined for each kind of query performed by the system. Also because of that, this evaluation must be executed together with query processing, in which the Quality Evaluation Module interacts with the Query Processing Module as explained in the next section.

Task-independent evaluation might also provide a global scores for a series. This option is not recommended, since queries for crop forecasts usually do not use an entire series. It is not coherent to provide the user with an indication of the quality of an entire series in queries that utilize only fractions of it. In the same way, a global metric that combines in one only score the evaluation of the four quality indicators was discarded. This approach is not adequate since the importance of each indicator varies with users.

C) Query Processing Module

After quality evaluation, the data are ready to be used by the Query Processing Module. This module has been partially implemented. A typical example of a query is the visualization, under graphical form, of statistics for each station, e.g., the rain distribution along the months of the year. Other results will require using a GIS to provide cartographic visualization, e.g., of rain spatial distribution.

The most important product to be provided is pluviometric forecasts, guiding the crop schedule of each geographic region. The system will also support prediction of climatic events prejudicial to agriculture, such as extreme rains, frosts and little summers, allowing the preventive actions.

The system also provides data on quality evaluation. This is achieved through an interaction between Quality Evaluation and Query Processing Modules, as shown in Figure 7. After the query is posed by the user, the Query Processing Module notifies the Quality Evaluation Module of the query specification. The Quality Evaluation Module consolidates the scores of the completeness and free-of-error task-independent indicators to the data used in the query, generating one single score for each indicator. Next, it performs task-dependent evaluation, generating scores to timeliness and appropriate amount of data indicators. The four scores are passed to the Query Processing Module, which presents them to the user, jointly with the query result.

By analysing the results, the user can refine his query, being able to test the query with other data sets until getting a result with the desired quality level. Moreover, at the end, the user is provided with credibility parameters of the obtained result, having more support for decision taking.

Several *views* were implemented to speed up data ac-

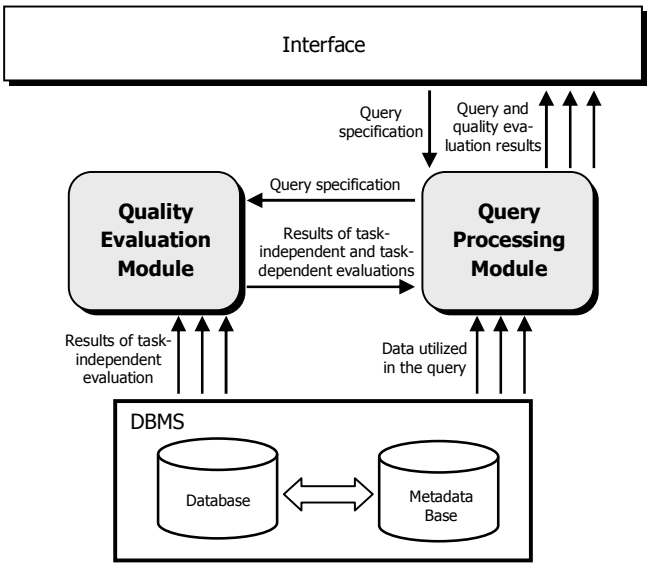


Figure 7: *Data flow in a query processing.*

cess, and for visibility purposes. The Query Processing Module uses them to access the integrated database, assuring that only data approved in the consistency phase are used in queries. Some views provide consolidated information about each historical series. Other aid data management, providing the data administrator with statistics about stations, like summaries of data quality assessments, and indicate, for example, the need for missing data estimation.

For performance reasons, the system explicitly stores annual and monthly statistics for each station, rather than computing them on the fly. The fact of dealing with more than 15,000,000 insertions disallowed a *on-line* processing of such statistics, which had to be implemented through batch processes.

4 Conclusions and current work

This text discussed an ongoing project which aims to develop a crop forecast support system, centered in pluviometric data integration. This system is being developed at CNPTIA-Embrapa, in cooperation with Institute of Computing of Unicamp. A great part of the system is already implemented and available in [3].

The main contributions of this article were the description of real problems of heterogeneous data integration in the context of spatio-temporal pluviometric data, and the proposal of a solution which combines spatio-temporal database concepts and geostatistical data processing. This solution takes data quality in consideration, which is a factor not usually available in real systems. Metadata help data access, and data quality evaluation enhance the support to decision taking.

The user will be able to perform queries on a database that integrates data from distinct sources and with different spatio-temporal granularities. As new institutions progressively join the system, it will be possible to estimate more missing data, increasing the quality of the database. This requires a procedure for "updating the past", one of the spatio-temporal characteristics of the system.

Products like temporal and geographical rain distributions, maximum rain forecasts and pluviometric behaviour identification will be available to various agricultural activity fields, helping experts in their tasks. Other areas like hydrology and urban construction will also be able to obtain benefits from the system.

In summary, the advances brought by the solution are: (a) a framework for integrating heterogeneous data, with the option of publishing wrapped data in XML using a well-defined schema; (b) easy access to integrated data through use of metadata; (c) definition of historical series consistency methods, including automated data re-evalution with the arrival of new data; (d) definition of a quality evaluation framework, whose results improve decision takings and (e) more precise forecasts and climatic behaviours identifications.

Current work involves implementation of query functionalities and automation of quality evaluation.

Acknowledgments

This work has been partially financed by Embrapa, CAPES, CNPQ, the Institute of Computing of Unicamp, the PRONEX-MCT SAI project and the WebMaps project from CNPQ.

References

[1] *Dublin Core Metadata Initiative*. `http://purl.org/DC` (accessed in August 2003).

[2] *World Meteorological Organization*. `http://www.wmo.ch` (accessed in August 2003).

[3] Embrapa Informática Agropecuária. *Agritempo - Sistema de Monitoramento Agrometeorológico*. `http://www.agritempo.gov.br`.

[4] G. Aslan and D. Leod. Semantic heterogeneity resolution in federated databases by metadata implantation and stepwise evolution. *The VLDB Journal*, 8(2):120–132, 1999.

[5] E. D. Assad. *Chuva no Cerrado - análise e espacialização*. Embrapa Cerrados, Brazil, 2001.

[6] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, pages 34–43, May 2001.

[7] N. R. Chrisman. Living with error in geographic data: Truth and responsibility. In *Proceedings GIS*, volume 1, pages 12–17, Vancouver BC, Canada, 1995.

[8] Open GIS Consortium. *Geography Markup Language (GML) Implementation Specification*. http://www.opengis.org/docs/02-023r4.pdf.

[9] D. Fallside. *XML-Schema Part 0: Primer*. http://www.w3.org/TR/2001/REC-xmlschema-0-20010502 (accessed in August 2003).

[10] US Government Federal Geographic Data Committee. *Content Standards for Digital Geospatial Metadata*. http://www.fgdc.gov/metadata/contstan.html (accessed in August 2003).

[11] D. Fensel. Ontology-Based Knowledge Management. *IEEE Computer*, 35(11):56–59, 2002.

[12] D. Fensel and M. A. Musen. The semantic web: a brain for humankind. *IEEE Intelligent Systems*, 16(2):24–25, 2001.

[13] R. Fileto. Issues on interoperability and integration of heterogeneous geographical data. In *III Brazilian Symposium on Geoinformatics - GEOINFO*, pages 133–140, Rio de Janeiro, Brazil, 2001.

[14] Food and Agriculture Organization of the United Nations. *The Agricultural Metadata Standards Initiative*. http://www.fao.org/agris/MagazineArchive/MetaData/TaskForceonDCMI.htm (accessed in August 2003).

[15] P. Hohl. *GIS Data conversion; Strategies, Techniques and Management*. Onword Press, 1998.

[16] J. Lima. *XML Schema for climatological data and metadata*. http://www.ic.unicamp.br/proj-adb/climatological/climatologicalSchema#.

[17] J. Lima. Integração de dados climatológicos heterogêneos. Master's thesis, Instituto de Computação, UNICAMP, Campinas, Brazil, 2003. To appear.

[18] D. S. Linthicum. Remember the metadata. *eAI Journal*, (9):8–10, September 2002.

[19] C. López, E. González, and J. Goyret. Análisis por componentes principales de datos pluviométricos. a) aplicación a la detección de datos anómalos. *Estadística (Journal of the Inter-American Statistical Institute)*, (46):25–54, 1994.

[20] C. López, J. F. González, and R. Curbelo. Análisis por componentes principales de datos pluviométricos. b) aplicación a la eliminación de ausencias. *Estadística (Journal of the Inter-American Statistical Institute)*, (46):55–83, 1994.

[21] M. Mecella, M. Scannapieco, A. Virgillito, R. Baldoni, T. Catarci, and C. Batini. Managing data quality in cooperative information systems. *10º Italian Symposium on Advanced Database Systems - SEBD 2002*, 2002.

[22] C. B. Medeiros. Bancos de dados espaço-temporais: Fundamentos e aplicações. In *VI Escola Regional de Informática - Anais*, pages 241–255, ICMC-USP, São Carlos, Brazil, 2001.

[23] C. B. Medeiros and A. C. de Alencar. Qualidade dos dados e interoperabilidade em SIG. In *I Workshop Brasileiro de Geoinformática - GEOINFO*, pages 45–49, Campinas, Brazil, 1999.

[24] M. Missikoff, R. Navigli, and P. Velardi. Integrated approach to web ontology learning and engineering. *IEEE Computer*, 35(11):60–63, 2002.

[25] I. Onyancha, F. Ward, F. Fisseha, K. Caprazli, S. Anibaldi, K. Johannes, and S. Katz. Metadata framework for resource discovery of agricultural information. In *Open Archives Inititiative Workshop, 5th European Conference on Research and Advanced Technology for Digital Libraries*, Darmstadt, Germany, 2001.

[26] World Meteorological Organization. *WMO Core Metadata Standard*. http://www.wmo.ch/web/www/metadata/WMO-core-metadata-toc.html (accessed in August 2003).

[27] L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[28] H. A. Rocha. Metadados para workflows científicos no apoio ao planejamento ambiental. Master's thesis, Instituto de Computação, UNICAMP, Campinas, Brazil, 2003.

[29] C. E. M. Tucci, editor. *Hidrologia: ciência e aplicação*, volume 4 of *Coleção ABRH de Recursos Hídricos*. Editora da Universidade, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, 1997.

[30] F. Wang and S. Jusoh. Integrating multiple web-based geographic information systems. *IEEE Multimedia*, 6(1):49–61, 1999.