Medical Informatics Europe '99
P. Kokol et al. (Eds.)
IOS Press, 1999

Application of the medical data warehousing architecture Epidware to epidemiological follow-up : data extraction and transformation

Elmostafa Kerkri[a], Catherine Quantin[a], Kokou Yetongnon[b], FA Allaert[a,c] and Liliane Dusserre[a]

[a]*Dijon University Hospital, medical informatics Department, Dijon, France*
BP 1542- 21034 Dijon Cedex- France- emkerkri@u-bourgogne.fr
[b]*University of Burgundy, Electronic and Informatic Engineering Laboratory, Dijon, France*
[c]*W2 "data security" European Federation of Medical Informatics*

Abstract. In this paper, we present an application of EPIDWARE, medical data warehousing architecture, to our epidemiological follow-up project. The aim of this project is to extract and regroup information from various information systems for epidemiological studies. We give a description of the requirements of the epidemiological follow-up project such as anonymity of medical data information and data file linkage procedure. We introduce the concept of Data Warehousing Architecture. The particularities of data extraction and transformation are presented and discussed.

Key words. data warehousing; epidemiological follow-up; data extraction and transformation; medical information systems; data anonymity and linkage

1 Introduction

Computer science has been used in medical applications since the 1970's. Health care systems in industrialised nations have undergone hospital output evaluation based on a classification of hospital stays called Diagnosis Related Groups (DRGs) aimed at curbing hospital expenditure. In 1985, the implementation of the application of the information system program (Programme de Médicalisation des systèmes d'information - PMSI) began in France. In 1996, the PMSI was enlarged to private medical sector and the Health ministry initiated the national medical network project Réseau Santé Social. The aim of this project is to propose an Intranet to enable medical professionals either to send information to sickness benefits organisms, or to exchange medical information between themselves.

This evolution in the application of computer science to healthcare followed the evolution in informatics, characterised by the transition towards microcomputers, towards networks and the Internet, and towards relational or object oriented databases.

Since 1992, Data warehouse technologies are used as components of decision support environment aimed at collecting and organising relevant information to aid decision-makers. Data warehouse technology can also be used in medical management. Besides this economic aim, we suggest using this technology to link recorded medical information concerning a given pathology in a given region in view of epidemiological studies. These

studies would also allow the optimisation of health expenses by limiting redundancies in examinations and procedures for the same patient while improving patient care. However, in order to respect the European legislation concerning personal data processing, and so to guarantee confidentiality, integrity and availability of data, this regrouping of information should be composed of different steps such as anonymity of the data, secured file transfer and linkage [3].

In the first part of this paper we present the "Epidemiological Follow-up" project and the necessary stages for its implementation. The second part is devoted to the data warehousing EPIDWARE architecture.

2 Development of the architecture EPIDWARE for epidemiological follow-up

2.1 The epidemiological follow-up project

To carry out epidemiological follow-up of diseases such as cardiovascular diseases or cancers, for example, it is necessary to link information recorded by different health structures participating in the care of a given patient. Diagnoses made by the Anatomo-pathology laboratory, type of the surgical operation and of chemotherapy treatments, the localisation and the duration of the different hospitalisations, and if possible the date of death.

To respect the European legislation concerning medical file linkage, we have developed an anonymous record linkage procedure [7,8,9], which is composed of two steps: an anonymity procedure and the record linkage of rendered anonymous files. The question now is to integrate the different files coming from different sources into a decisional database (data warehouse) while ensuring the regrouping of the information concerning a given patient.

2.2 Data warehousing concept

In the databases technology area, Data warehouse is defined as a collection of decision support functionalities / environment to allow decision-makers to take more rapid and relevant decisions. The Data warehouse has to insure the coherence of information of the enterprise and to facilitate its access. These objectives imply the integration of request tools and other tools for information analysis and presentation. Data have to be carefully gathered from different information sources, then cleaned and filtered and to be distributed only after validation of their quality. Thus, information provided by data warehouse is not a crude data but an information that assist and helps the decision-makers. It is subject oriented, aggregated, easily accessible, reliable, relevant, non-volatile and possesses a specific temporal context [1,2,4,5].

The principle of data warehousing is to extract useful information, to combine and to consolidate them into a coherent data repository to provide a single image of activity reality. This improves data quality and allows users to retrieve necessary data by themselves [6]. The coherence of data is measured globally, based on the viewpoint of the manager who seeks for example if data is complete and without contradiction [5].

2.3 EPIDWARE architecture

The implementation process of a data warehouse is generally progressive. Often, it begins with the implementation of a data mart (a data warehouse relative to a given

department/activity). Then, two types of development are possible according to the organisational choice of the administrative staff of enterprise: progressive centralisation of strategic data or implementation of as much data marts as specific services.

The modelling of the target data structure (a data mart or a data warehouse) constitutes the initial stage of data warehousing. Builders establish, with the direction of the enterprise, a dictionary of data description (metadata). They also specify extraction, translation and integration tools. Once implemented, the data warehouse has to be updated to ensure that the format of data corresponds to the user's need, which may change later on. More, changes in sources of information have to be propagated on the data warehouse.

Once extracted, data are aggregated and filtered so as to harmonise their formats and to eliminate redundancies. The attribution of coherent values to non initialised variables (, completes the stages which precede the information loading in the Data warehouse.

Information sources: Information sources (Figure1), can be multiple [1]: database systems, files systems, document HTML or knowledge bases. In medical context, information sources are in public or private hospitals, biomedical analysis laboratories, radiology departments. And data are related either to patient identity or to medical information.

Wrapper: Is mainly composed by a three elements: the Translator, the Monitor and the Anonymity tool. **Translator**'s role is to translate data from the format and the model of the information sources to the format and the model of the data warehouse. The translation consists in making the underlying information source appearing as subscribing to the data model used by the warehousing system. For example, if the information source has a data file structure but the warehouse uses a relational data model, then the wrapper/monitor has to have an interface presenting data of the information source as if they were relational[1].

The detection of data changes concerning the data warehouse and their propagation to the integrator is the main role of the **Monitor**. Then, the changes occurring on data have to be translated by the wrapper, from the format and the model of the information source, to the format and model of the data warehouse, just as for data themselves. Another approach consists in ignoring the functionality of change detection and transmitting periodically whole data copies from the information source to the data warehouse. The Integrator can

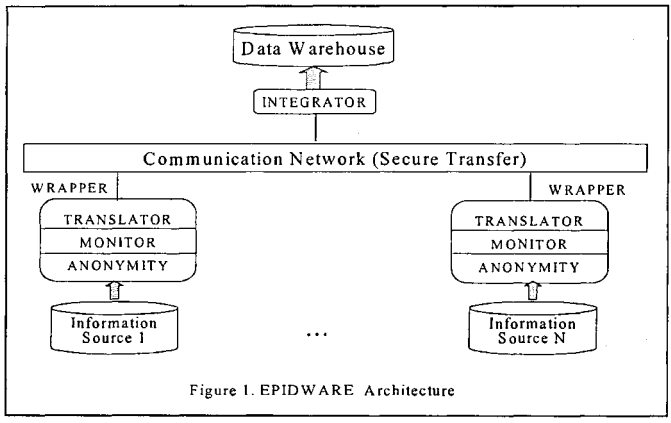

Figure 1. EPIDWARE Architecture

[1] Most of the commercial Data Warehousing systems suppose that the information sources and the Data Warehouse are relational.

then combine these data with those of the data warehouse coming from other sources, or ask all sources for other information to complete data of the warehouse. However, if the continuous access of updated information is required, it is preferable to adopt the principle of change detection and their propagation to data of the warehouse. The first stage of **Anonymity** procedure is included in the Wrapper functionalities.

Integrator: To load information in the data warehouse, the integrator must filter, refine, integrate and combine information coming from the local sources. It provides decision-maker with a clear, relevant and non-redundant view of patient medical information. To achieve the above, the integrator include some additional functionalities. First, it receives change notifications from the wrapper and updates the data warehouse. Second, it applies the second stage of anonymity to the received files. Finally, it links received medical files to regroup medical information of each patient.

2.4 Data extraction

The particularity of our project is that most administrative and medical information sources are outside our hospital. Usually, we do not meet any difficulty with the database systems bought from computer companies specialised in medical fields. But some hospitals use software without any technical documentation and it is thus difficult to locate, for example, the database tables of discharge abstracts containing associate diagnoses and the corresponding acts. Moreover, sometimes information needed for epidemiological studies is missing, for example the term of the pregnancy for perinatology studies, which then require a completion of the data set manually.

Another problem is to eliminate duplicated data, due to a lack of quality of administrative information. In some cases we have data redundancy characterised by as many records, for each hospital stay of the patient, as numbers of therapeutic procedures and associate diagnoses coded for this stay. In order to solve the problem of regrouping all these records in a single one, we have implemented a solution, consisting of three steps. The first step is to eliminate redundancy by therapeutic procedure, the second step consists of doing the same thing with the associate diagnoses, and the third, of joining the two files resulting from the two first steps. This solution was implemented with Visual Basic and SQL for Microsoft Query.

In a particular application of neonatalogy follow-up, the problem of the connection between the record of a baby and his mother was posed. If the mother and her babe do not have the same name, we need to use our linkage procedure [8,9] to know the medical information about the mother. In practice, giving successive identification numbers to the baby and his mother enables to link the babe and his mother records. This seemed to solve the problem, but not completely because we have to pay attention to the case of twins for example. We proposed a general solution by creating a new variable, namely the identification number of the mother, in the baby's record.

2.5 Data transformation

Data are translated according to the metadata specifications. For example, date of birth must be coded as DDMMYYYY, so we have to translate it from any other format to this one. In other way, date of birth is indirectly nominative, thus the second translation is to compute the age of the patient from his date of birth. The zip code, which is also indirectly nominative, must be replaced by a geographical code.

The most important transformation concerns nominative variables such as the name, the first name and the maiden name. The first stage consists in spelling treatment [9] that

contains rules and facts. This step is necessary because of typing errors, in particular for homonyms. Obviously, the knowledge base of the spelling treatment must be adapted to the language used. The second stage consists in an irreversible transformation of all nominative variables.

3 Discussion and conclusion

The advantage of EPIDWARE information system is to combine data warehouse techniques with an anonymity procedure to design an epidemiological follow-up system. It is very useful, from an epidemiological point of view to be able to integrate and to consolidate inconsistent medical data from various legacy systems into one coherent data set and to improve data quality by extracting relevant and pertinent information. We cannot ignore the difficulties of obtaining information, which may even imply manual copying from existing databases. Indeed, the data-processing solutions proposed depend on the concerned information sources' specificity. Though, these solutions will progressively constitute a data extraction toolkit well adapted to most information sources so that manual intervention would be exceptional. In the same idea, the set of transformation roles will be completed in order to take into account new information sources particularities.

Epidemiologists can benefit also from direct access to the information needed without requiring the help of computer scientists, using requests on databases already implemented. Of course, this necessitates a certain sense of anticipation. At last, this architecture relieves epidemiologists from concerns about the respect of legal security aspects as it guarantees data confidentiality and secures information processing.

Acknowledgements. This work was supported by the Burgundy Regional Council, France Telecom and the French "Ligue Bourguignonne" against cancer.

References

[1]]Widom J, Research problems in Data Warehousing, Proc. Of 4th Int'l CKIM, Nov. 1995.
[2] Franco JM & EDS- Institut prometétéus, Le Data Warehousing, ed Eyrolles, Paris, janvier 1997
[3] Quantin C., Kerkri E. Allaert FA, Bouzelat H, Dusserre L. Security aspects of medicale file regrouping for the epidemiological follow-up. In Cesnim B, McCray A, Scherrer JR (eds), Proceeding Medinfo'98, pp. 1135-1137
[4] Inmon WH & Hackthorn RD. Using the data warehouse, Wiley-QED Publication, 1994
[5] Kimball R. The Data Warehouse Toolkit.(French version by Raimond C.) Paris: International Thomson Publishing France, 1997.
[6] Sakaguchi T. and Frolik Mark N. A Review of the Data Warehousing Literature. Web: http://www.people.memphis.edu/~tsakagch/dw-web.htm. Jan. 31, 1996.
[7] Bouzelat H, Quantin C, Duserre L. Extraction and Anonymity. Protocol of Medical file. JAMIA 1996 Symposium Supplement:323-7.
[8] Quantin C., Bouzelat H, Allaert FA, Benhamiche AM, Faivre J, Dusserre L. How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure. Int J of Medical Informatics, 49 (1998) 117-122.
[9] Quantin C., Bouzelat H, Allaert FA, Benhamiche AM, Faivre J, Dusserre Development of an automatic record hash coding and linkage procedure to warrant epidemiological follow-up data confidentiality. Methods of Information in Medicine. 1998; 37:271-277.