Heterogeneous Database Interoperability Using the WWW*

Gustavo Zanini Kantorski
Cora H. F. Pinto Ribeiro
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Av. Bento Gonçalves, 9500 – Bloco IV – Agronomia – Caixa Postal 15064
91501-970 – Porto Alegre – RS – Brasil
Fax +55(51)336 5576
e-mail: [gustavoz, cora]@inf.ufrgs.br

Abstract

The integrated access to distributed heterogeneous information is a major concern in large corporations and governmental institutions. However, integrated access to disperse information should not require the eradication of local systems heterogeneity and autonomy because of the financial impact associated to such solution. The use of available Internet tools turns out as an affordable, simple, and also platform and DBMS independent, alternative for integrated information retrieval of heterogeneous databases. This paper presents a tool for integrated access to heterogeneous databases through the Internet. The proposed solution is supported by the previous mapping of all participant conceptual schemata and by the use of available Internet technology.

***Key words**: conceptual schema mapping, heterogeneous databases, database interoperability, Internet, information retrieval.*

1 Introduction

The integration of data from heterogeneous databases remains as an unsolved problem in the database community. Existing heterogeneity results from hardware and software technological evolution, distribution of data inside corporations, data legacy, and diversity of problem oriented applications.

Most of the available techniques for integrated access to heterogeneous databases are supported by the previous conceptual models integration of participant databases [1, 2], or by the inclusion of an additional layer of software for logical data integration [3]. The first approach is based on the creation of an integrated global conceptual schema, after the conforming of local schemas to the global model. The integration process may call for troublesome, expensive, and time-consuming modifications on local conceptual model and applications [4]. Besides, this may bring restrictions to local autonomy. The second approach is usually implemented by frameworks, which integrate specific data sources, and use a particular query language, which must be learned by users.

Although integrated access to available information from heterogeneous data sources remains with no ultimate solution, access to distributed data is easily implemented by using free WWW available resources, without any restriction on local hardware and software.

* This work was partially supported by the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul – FAPERGS, under Grant #96/1074.6.

This work presents a tool for integrated access to heterogeneous database through the Internet, supported by the previous mapping of participant databases conceptual model. The development of such tool was based on two assumptions: a) the use of WWW resources for database access, b) integrated access to available information without any conceptual schema, database or local applications modifications.

The next section reviews concepts and related work on heterogeneous databases. Section 3 describes the conceptual schema mapping, proposed by Ribeiro [5], and used in this tool. The tool for integrated access to heterogeneous databases is described in Section 4. Conclusions and future directions are discussed in Section 5.

2 Background and Related Work

A wide range of approaches and methodologies have been proposed for the integrated access of information from autonomous databases, stored in different software and hardware platforms, including federated database systems [2], multidatabase systems [6], heterogeneous database systems [7], and mediators [3]. The above methodologies adopt a different viewpoint regarding the use of a global conceptual schema, which defines the common and sharable part of all participant databases, with no redundancies [8].

The creation of the global conceptual model is supported by local conceptual model comparison, equivalence identification, conflict identification, and conflict resolution. The last step is usually associated to the conforming of local representation of data to global standards, with modifications on local data structure and related applications [9].

Solutions based on mediators [3] provide data access through a specific software layer between users applications and data sources. This middle-layer provides data access and data integration to users. However, this solution requires additional processing and the inclusion of rules for each participant database, based on the previous integration of local schemas.

The bellow methodologies represent the state of the art in integrated access to heterogeneous sources. Though all of them provide access to different kind of information sources[1], the following text analyses the handling of structured data sources, according to the focus of the present work. This restriction is due to the fact that this tool was designed to support integrated query of heterogeneous databases.

Singh [10] presents a Tesserae Integration Engine - TIE for distributed information access. TIE comprises the description of the data, referred as metadata, which is used to infer relationships among objects and to unify representations of heterogeneous data into a common model. This methodology performs the integration of information according to the data sources metadata specifications, to the user's query models and to the logic of business rules. The metadata is also used, during the query processing, to decompose complex requests into simple requests, to route requests for the appropriate sources and to integrate the results. The information query is implemented through HTML pages, in a standardized way, allowing the visualization of the shared data and of the local data source.

Subrahmanian et al. [11] developed the Heterogeneous Reasoning and Mediator System - HERMES, based on the Hybrid Knowledge Bases - HKB [LU 94] theory. This theory is based on a declarative language, used on the definition of mediators, which expresses the semantic integration of information from several sources of data. HERMES uses mediators as a starting point for the integration of information from heterogeneous sources. However, the system requires a specific implementation for each supported database, like

[1] Information sources are classified into three different classes: structured, semistructured and non-structured. Databases are structured information sources. Semistructured sources include most of the information found on the WWW, with known data fields. Non-structured sources include text files, images and other WWW pages.

DBASE, INGRES and PARADOX. The query language adopted in the Hermes system is a logical rule-based language. For query purposes, explicit database domain knowledge may be required. The HERMES query interfaces are platform dependent, but there is also a Web interface, based on CGI scripts.

Chawathe et al. [12] presents The Stanford-IBM Manager of Multiple Information Sources - TSIMMIS project [13, 14], for the integration of heterogeneous structured and not structured data sources. The TSIMMIS architecture is sustained by the use of mediators and wrappers for data integration. The TSIMMIS provides integrated access to information by using a common data model named Object Exchange Model - OEM. The basic part of the OEM, called object, is composed of four elements: label, for naming and semantics information; type, for object type identification; value, for literal or nested objects enumeration; and object-id, for unique object identification (for each data source). A visualization tool was developed for query through the WWW. However, in order to formulate a query, users should understand, and use, a specific query language, called OEM-QL.

The above works support the heterogeneity of different types of information sources. Still, the integrated access to heterogeneous structured data, focused on this paper, presents limitations on information access and query, and on usability. The visualization of the information in HERMES and TSIMMIS is limited to the common parcel, in a uniform and transparent way, without allowing the visualization of the available additional information locally. TIE allows the visualization of additional information by means of a navigational tool. However, all additional information interfaces are not standardized

The user perception of all available information as from one unique database makes it impossible to distinguish specific data sources. This solution may be an important restriction in the medical field, because it avoids the identification and mapping between data and data source (for example, establishing a relationship between the final diagnostic and the results of a particular diagnostic test). In such environment, the possibility of data grouping according to data source is an important requirement.

Solutions that use SQL for query narrow down the group of users able to interact with the system.

The methods discussed above handle all sources of information. However, access to different sources of heterogeneous information requires complex interfaces. Such interfaces are hard to develop, to use and to update. Besides, if the methodology does not provide the matching of all data referring to the same real life entity, the integration of different data sources may be ineffective.

3 Conceptual Schema Mapping

Conceptual schema mapping provides an alternative to heterogeneous databases integrated access, without the requirement of local schema integration. This paradigm [9] allows a global and complementary visualization of all available information of a same real world entity. Different parcels of the universe of discourse, represented in different databases, are put all together, without the burden of representation reconciling. Such approach supports the representation of multiple views, still providing an integrated global view without the requirement for discrepancies reconciliation. This solution combines multidatabase (maintenance of local database autonomy) and federated databases (creation of a global conceptual schema) principles. In this methodology, the global conceptual schema is replaced by a main structure (central registry) were all information about participant databases, as well as mapping and equivalencies among heterogeneous representations, is stored.

Each participant database specifies all local sharable information through the creation of an export schema – ES. Each ES is represented in a canonical model. The use of a unique canonical model, as adopted by traditional database integration methodologies [8, 15, 16, 17], eliminates data model heterogeneity. The ES is semantically enriched with the inclusion of additional information about local attributes type and domain. In addition, all real life entities represented in the ES are listed in a table of local objects[2] – TLO, including a textual description of each. The ES and TLO support the equivalence and conflict identification process, in order to establish equivalences among heterogeneous databases representations. All ES and TOL, as well as other methodology elements, are stored in the central registry – CR. The other components stored in the CR are: table of objects equivalency – TOE, comprising information about equivalent real life objects stored in different EE; table of attributes equivalency – TAE, including information to support object identification, attributes equivalence, and attributes domain mapping. The TAE may also include procedures for domain conversion and discrepancy report about equivalent attributes instantiation conflicts (to be detected in the query process).

The mapping process [9] includes two steps:

1. ES comparison, after the identification of TLO equivalent objects, for modeling equivalencies and conflict identification;
2. TOE and TAE creation, including detected similarities, conflicts and differences of representations, and procedures for domain conversion and objects instantiation discrepancy report.

The mapping methodology provides integrated access to information from heterogeneous participant databases putting together all locally represented attributes of a same real life object. This visualization occurs in an integrated way, but it is possible to associate information to data source through the interface (information is grouped according to the original source database). In large organizations, participant databases are often independently owned (for example, database system integrated to a medical equipment). This situation makes any database update or modification unacceptable. However, discrepancies on instantiated equivalent data may be critical and the detection of instantiated attributes can be very helpful [4]. The mapping methodology deals with instantiation conflicts through the inclusion of warning procedures, without any local autonomy restriction. The conflict detection occurs during the query process.

4 The proposed tool

The tool presented in this paper provides a wide integrated access to heterogeneous databases, through the WWW. Information search and visualization is performed after the mapping of participant databases, as proposed by Ribeiro [9].

The generation of components ES, TLO, TOE, and TAE, proposed by Frederes [18], is out of the boundaries of this paper.

4.1 Tool Components

All the components of the tool are stored in the central registry (Figure 1). The central registry is implemented as a database, were all metainformation about participant databases, including exported information, data dictionary, data equivalencies, and data mapping, is stored as tables. Information stored in CR supports the entire query performed by users. The tool is not limited to a predefined range of databases, and new participant databases may be always included. The inclusion of new participant database does not require any modification

[2] In the present work, object refers to a real life entity, represented in different heterogeneous databases.

on the software, because the integrated access is dynamically performed, following the metainformation stored on the CR. Whenever CR information is modified by inclusions, exclusions or updates, the visualization of all available information is also updated.

The original mapping methodology components were modified by the inclusion:

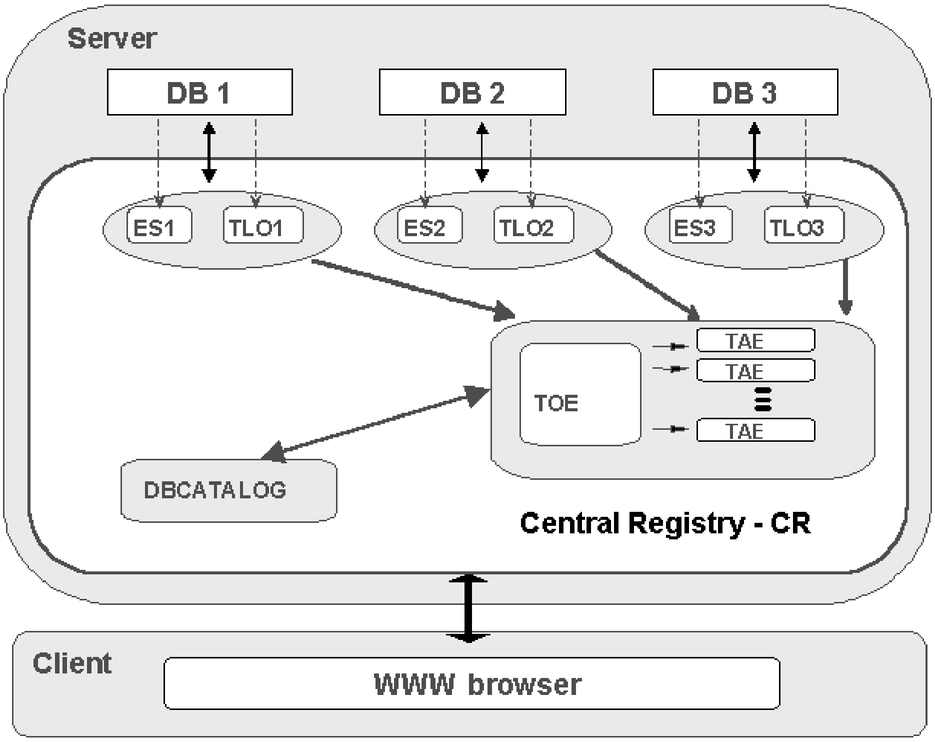


Figure 1: Tool Components (abbreviations as described in the text)

1. in the ES, of a text description for each exported attribute, to be used in the dynamic generation of the user interface;
2. of the DBCATALOG component, containing all necessary information, for access purposes, of participant databases, including database name, database connection alias, user connection, and JDBC driver name.

In the proposed tool, conflicting domains (for example, "M" and "F" versus "1" and "2", for gender representation) are visualized according to local standards.

When the user starts the tool, all available integrated information is shown in the screen, by means of a menu. This approach requires from the user no special or previous knowledge about any query language or about participant databases structure.

This tool does not support domain conversion procedures and object instantiation discrepancy report, included in the original mapping process proposed by Ribeiro. However, the inclusion of such components is already planned as future work.

4.2 Tool Architecture and technical issues

User interaction with the tool occurs via an HTML page and query is implemented by Java applets. This solution (Figure 2) makes query easier. In addition, distributes processing among clients and application servers. The reasons for the use of Java in the heterogeneous databases access are: the possibility of dynamic components creation, facilities for data recovery from different databases, execution of the tool at the client's process space, and platform independence. Such characteristics turn Java a feasible alternative for heterogeneous systems interoperability by use of the WWW.

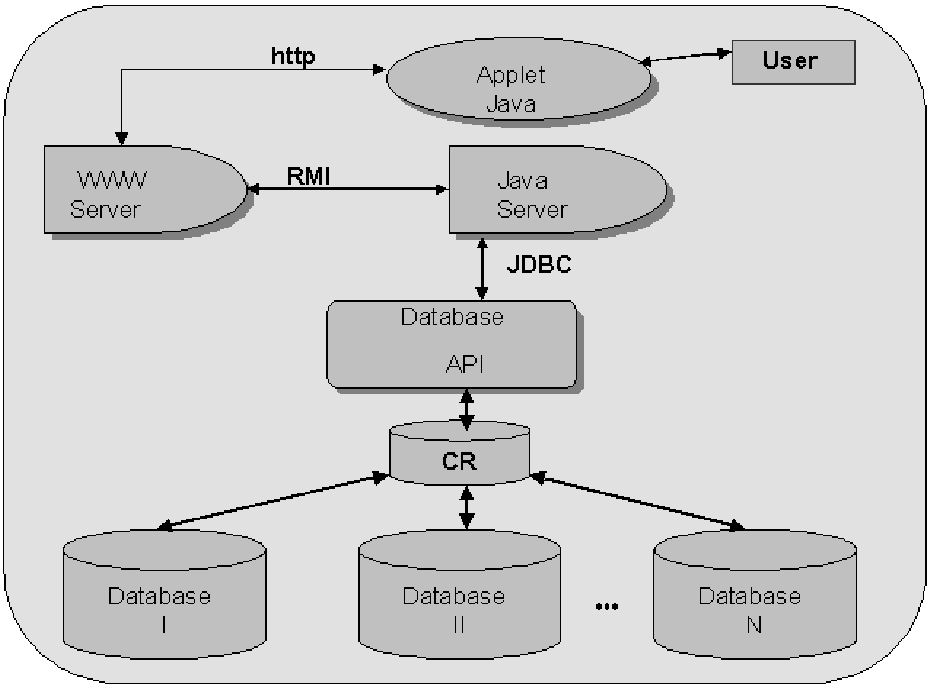
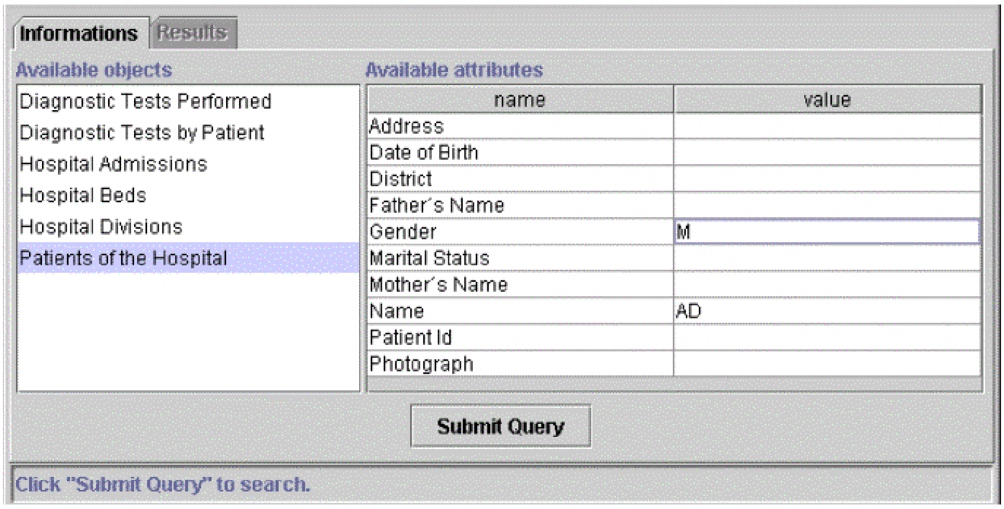


Figure 2: Tool Architecture

The standard user interface of the tool is dynamically created, based on the metainformation stored in the CR. The initial menu includes all objects informed in the TLOs. After object selection, all matching available attributes are also visualized, based on the ESs description. Equivalences are identified after the TEO and TEA tables. Object and attribute equivalency is shown in an integrated complementary way. Query results are included in the HTML page by Java applets.

The tool allows user query of object and object attributes by pointing. The access to selected databases is then performed, based on user request. Query results are than shown on the screen.



Available objects

- Diagnostic Tests Performed
- Diagnostic Tests by Patient
- Hospital Admissions
- Hospital Beds
- Hospital Divisions
- Patients of the Hospital

Available attributes

name	value
Address	
Date of Birth	
District	
Father´s Name	
Gender	M
Marital Status	
Mother´s Name	
Name	AD
Patient Id	
Photograph	

Click "Submit Query" to search.

Figure 3: Query Interface (Java applet)

The initial interface of the tool is divided into two sections (Figure 3).

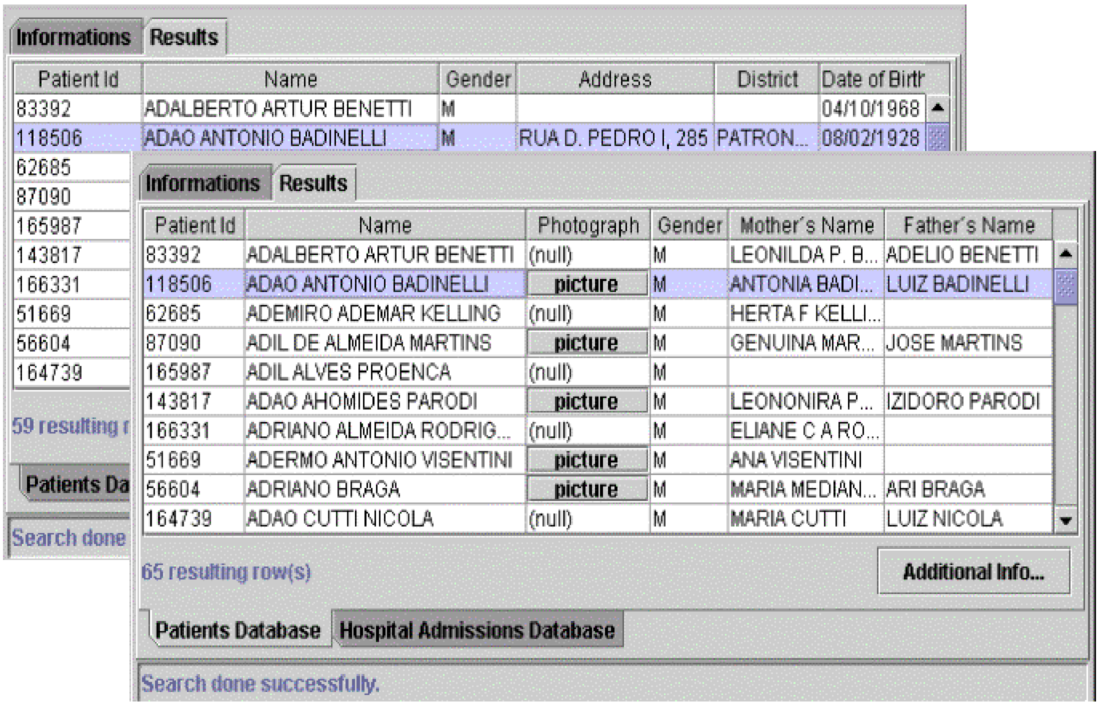
The first one, called *Available objects*, allows user to select one of the available objects from participant databases. Every available object is reported once, even if it is represented in more than one database. At this point, the relationship between object and database is transparent to users.

The second component of the initial interface, called "*Available attributes*", allows the specification of constraints for object selection. After object selection, attributes from all ES representation of this object are listed. Again, equivalent attributes from different ES are listed once.

The SQL expressions are then generated, based on user selection and ES information, in order to get selected information from selected databases. Attributes used as selection criterion are translated into selection filters. Access to different databases is performed based on information stored in the DBCATALOG component of the CR.

Query results are visualized after the recovery of all related data from different databases. Still, resulting data is grouped according to specific data source (Figure 4), allowing the mapping of data source and query results. Whenever new participant data base information is included in the RC, the tool automatically generates a new interface component, identified after the database name.

The tool also allows the visualization of additional information about the object (Figure 5), stored as attributes of other real life objects, but related to it (as a diagnostic test performed on the patient, or hospital admission information about him).



Patient Id	Name	Gender	Address	District	Date of Birtr
83392	ADALBERTO ARTUR BENETTI	M			04/10/1968
118506	ADAO ANTONIO BADINELLI	M	RUA D. PEDRO I, 285	PATRON...	08/02/1928
62685					
87090					
165987					
143817					
166331					
51669					
56604					
164739					

59 resulting r

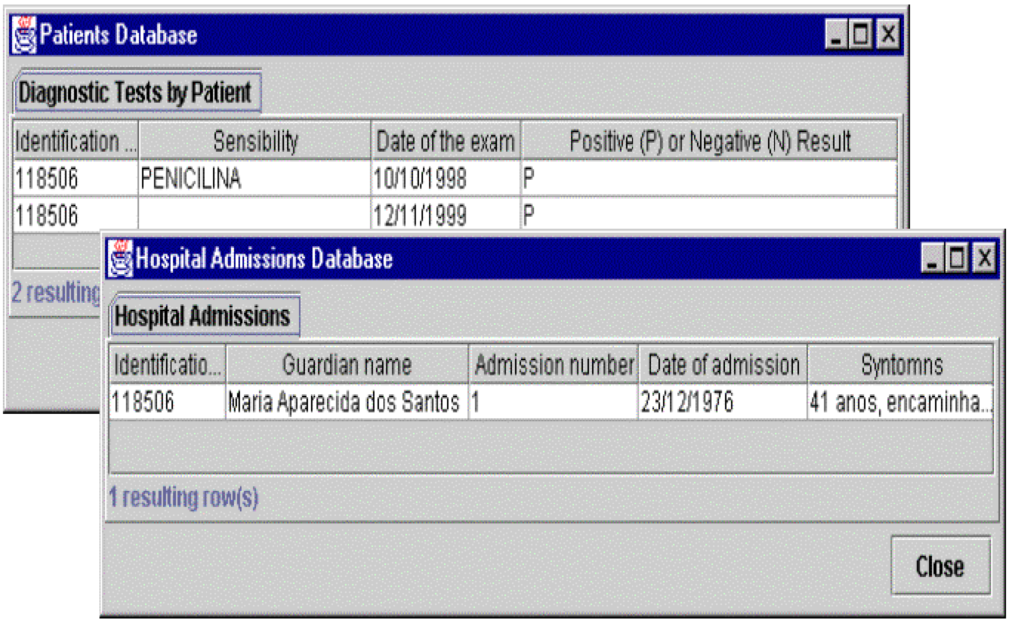
Search done

Patient Id	Name	Photograph	Gender	Mother´s Name	Father´s Name
83392	ADALBERTO ARTUR BENETTI	(null)	M	LEONILDA P. B...	ADELIO BENETTI
118506	ADAO ANTONIO BADINELLI	picture	M	ANTONIA BADI...	LUIZ BADINELLI
62685	ADEMIRO ADEMAR KELLING	(null)	M	HERTA F KELLI...	
87090	ADIL DE ALMEIDA MARTINS	picture	M	GENUINA MAR...	JOSE MARTINS
165987	ADIL ALVES PROENCA	(null)	M		
143817	ADAO AHOMIDES PARODI	picture	M	LEONONIRA P...	IZIDORO PARODI
166331	ADRIANO ALMEIDA RODRIG...	(null)	M	ELIANE C A RO...	
51669	ADERMO ANTONIO VISENTINI	picture	M	ANA VISENTINI	
56604	ADRIANO BRAGA	picture	M	MARIA MEDIAN...	ARI BRAGA
164739	ADAO CUTTI NICOLA	(null)	M	MARIA CUTTI	LUIZ NICOLA

65 resulting row(s)

Search done successfully.

Figure 4: Query Result Interface



Identification ...	Sensibility	Date of the exam	Positive (P) or Negative (N) Result
118506	PENICILINA	10/10/1998	P
118506		12/11/1999	P

Identificatio...	Guardian name	Admission number	Date of admission	Syntomns
118506	Maria Aparecida dos Santos	1	23/12/1976	41 anos, encaminha...

1 resulting row(s)

Figure 5:Query Result - Additional Information

5 Conclusions

In large organizations, several independent databases systems are usually simultaneously used. The integrated access to these data is not an easy task, but the financial and time demanding aspects associated to modifications on existing systems turns such solution unacceptable The possibility of multiple and integrated visualization of a same real life object, from independent and autonomous data sources, matches the need for integrated retrieval of information units that are related, but exist in separate, autonomous and heterogeneous databases.

This paper presents a tool for integrated access to information from heterogeneous databases, through the WWW. The integrated access is based on the conceptual mapping of participant databases export schema, without the requirement of any local modification of database structure or applications.

This tool uses a friendly graphical interface for user interaction. This approach does not require the previous knowledge of any query language by the user. Also, the query interface is platform and DBMS independent. Query results may be visualized as specific database views, allowing the structured navigation throughout participant databases.

An important issue granted by the tool is the dynamic inclusion of new data sources: whenever a new database is included, all information stored in the CR at query time is automatically shown to user. In the same way, modification on or exclusions of participant databases are immediately indicated in the tool interface.

The aim of the tool is to provide integrated read-only access to heterogeneous databases. The reason for avoiding updates comes from the original real life application field for this tool: the medical field. In this environment, several independently owned heterogeneous databases, attached to special equipments, are simultaneously used. The acceptance to share information among existing databases is not extended to allowing modifications on locally owned data. At the same time, the development of a centralized system is unacceptable, because of the financial costs and time demanding aspects associated to this solution. Still, allowing the integrated access to available information is a major issue and provided the foundation for the present work.

The use of the Java language provides friendly user interfaces, powerful search engines, and compatibility with different platforms, browsers and databases. It also allows the execution of the tool at the client's process space and provides access to any database through the WWW.

Comparing to other methodologies [11, 10, 13, 19, 20], the present tool contributes with an easy to use powerful tool for integrated access to heterogeneous databases. The adopted solution provides access to a wide range of DBMS, without the requirement of specific middleware construction.

At the current point, this tool is being used at the Federal University of Santa Maria (UFSM) teaching hospital, integrating medical data from DB2, ACCESS and Oracle databases. The tool is also in use at the Data Processing Center of the UFSM, for comparing existing databases structures and contents.

Warning procedures for conflicting instantiated data have not been introduced yet, and the detection of data discrepancy remains visual. The tool usability may be improved with the inclusion of data conflict alert components, for data integrity purposes: even without performing local updates, the detection and warning of differences among equivalent instantiated data may be extremely useful.

The possibility of suitable access, for local database input purposes, of available duplicated data from external databases, is also being considered as an improvement for the tool.

Pictures and sounds may be also stored in commercial databases as structured data. This tool may be easily improved for image and sound data sources access. However, query selection, using pictures and sounds as filters, has not been addressed yet.

References

[1] RAM, S. Heterogeneous Distributed Database Systems. *Computer,* New York, 24[1]: 7-9, December 1991.

[2] Sheth, A.; Larson. J. Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases. *ACM Computing Surveys,* New York, 22[3], September 1990.

[3] Wiederhold, G. Mediators in the architecture of future information systems. *IEEE Computer*, 25[]:38-49, 1992.

[4] Ribeiro, Cora H. F. P.; Oliveira , J.P.M.de. Multidabase Interoperability: An Approach on Extracting Semantic Equivalence, In *Proceedings of the Symposium on Software Technology* - SoST'99, Buenos Aires, Argentina, September 1999.

[5] Ribeiro, Cora H. F. P. *Heterogeneous Databases: Conceptual Schema Mapping in an Object Oriented Model.* PhD Dissertation. UFRGS, Porto Alegre, Brazil, 1995. (In Portuguese)

[6] Bright, M.; Hurson, A.; Pakzad, S. A taxonomy and current issues in multidatabase systems. *Computer,* New York, 25(3):50-62, 1992.

[7] Chung, C. DATAPLEX: an Access to Heterogeneous Distributed Databases. Communications of the ACM, New York, 33(1):70-80, January 1990.

[8] Batini, C.; Lenzerini, M.; Navathe, S. A Comparative Analysis of Methodologies for Database Schema Integration. *ACM Computing Surveys*, 18(4): 323-64, December 1986.

[9] Ribeiro, Cora H. F. P.; Oliveira , J.P.M.de. Multidatabase Interoperability Through Conceptual Schema Mapping in an Object Oriented Model, In *Proceedings of the IX*

International Conference on Parallel and Distributed Computing Systems (PDCS96), Dijon, France, September1996.

[10] Singh, N. Unifying Heterogeneous Information Models. *ACM Computing Surveys*, New York, 41[5]:37-44, May 1998.

[11] Subrahmanian, V. et al. *HERMES: A Heterogeneous Reasoning and Mediator System.* Available in http://www.cs.umd.edu/projects/hermes/overview/paper/index.html (April 1999).

[12] Chawathe, S.; Molina, H.G.; Hammer, J.et al. The TSIMMIS Project: Integration of Heterogeneous Information Sources. In *Proceedings of Anniversary Meeting of the Information Processing Society of Japan*, Tokyo, Japan, 1994.

[13] Hammer, J.; Aranha, J.; Ireland, K. *Browsing Object-Based Databases Through the Web.* Technical report, Stanford University, 1996.

[14] Garcia-Molina, H. et al. The TSIMMIS approach to mediation: Data models and Languages. *Journal of Intelligent Information Systems*, 1997.

[15] Ahmed, R. et al. The Pegasus Heterogeneous Multidatabase System. *Computer*, New York, 24(12):19-27, December 1991.

[16] Kaul, M.; Drosten, K.; Neuhold, E.J. Viewsystem: Integrating Heterogeneous Information Bases By Object-Oriented Views. In *Proceedings of IEEE International Conference Data Engineering,* Los Angeles, USA, February 1990.

[17] Reddy, M.P. et al. A Methodology for Integration of Heterogeneous Databases. *IEEE Transactions on Knowledge and Data Engineering,* New York, 6[6]: 920-933, December 1994.

[18] Frederes, S; Ribeiro,C.H.F.P; Oliveira , J.P.M.de. Ferramenta de Apoio a Conversão de Esquemas Conceituais Heterogêneos. In *Proceedings of the Symposium on Software Technology* - SoST'99, Buenos Aires, Argentina, September 1999. (In Portuguese)

[19] Vidal, V.M.P., Loscio, B.F. Especificação de Mediadores para Acesso e Atualização de Múltiplas Bases de Dados. In *Proceedings of the XII Simpósio Brasileiro de Bancos de Dados – SBBD'97, Fortaleza, CE, Brazil, October 1997.* (In Portuguese)

[20] Uchôa, E. M. A.; Lifschitz, S.; Melo, R.N. Interoperabilidade em um Sistema de Bancos de Dados Heterogêneos usando Padrão CORBA. In *Proceedings of the XII Simpósio Brasileiro de Bancos de Dados* – SBBD'97, Fortaleza, CE, Brazil, October 1997. (In Portuguese)