Data Engineering: Modeling and Integration Issues

Ana Carolina Salgado

▶ **To cite this version:**

Ana Carolina Salgado. Data Engineering: Modeling and Integration Issues. Computer Science [cs]. Université de Versailles-Saint Quentin en Yvelines, 2008. tel-00324525

HAL Id: tel-00324525
https://theses.hal.science/tel-00324525

Submitted on 25 Sep 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université de Versailles Saint-Quentin-en-Yvelines

Habilitation à Diriger des Recherches
Spécialité Informatique

Data Engineering: Modeling and Integration Issues

Mémoire présenté par

Ana Carolina Salgado

Le 26 Mai 2008

Devant le Jury

Rapporteurs	Aris Ouksel	Professeur – Université de l'Illinois, USA
	Stefano Spaccapietra	Professeur – Ecole Polytechnique Fédérale de Lausanne
	Patrick Valduriez	Directeur de Recherche – INRIA
Examinateurs	Zohra Bellahsène	Professeur – Université de Montpellier
	Mokrane Bouzeghoub	Professeur – Université de Versailles Saint-Quentin
	Patrick Brézillon	Directeur de Recherche - CNRS

A meus pais,
que me ensinaram o caminho

A meu filho,
que me dá forças para continuar

Acknowledgements

My first and main acknowledgement goes to Mokrane Bouzeghoub for having me as a member of the team in his research group and who encourages me to write this HDR. He also accepted my students as his, enriching their work with outstanding research directions. All my gratitude to his friendship and kindness.

A special thanks to the *rapporteurs*, Aris Ouksel, Stefano Spaccapietra and Patrick Valduriez, for their availability to read and contribute to the text. Also, Zohra Bellahsène and Patrick Brézillon who agreed to participate in the evaluating committee, my great thanks.

During the last year I was part of a group that accepted me as one of them. Many thanks to all members of SIAL group. It was really a pleasure to be among them during this period of time. Particularly, I would like to express my appreciation to Zoubida Kedad with whom I shared not only the office but also my research activities, my doubts, my problems,… Thank you very very much, Zoubida, for your friendship and availability. We had a great time together!

My thanks to all the staff of PRiSM.

This report could not be done without the work produced by ALL my students (PhD, MSc and undergraduate) that participated in my research projects. Particularly, I would like to nominate those whose research results are included in this report (in chronological order): Valéria Times, Ricardo Ciferri, Valéria Soares, Bernadette Lóscio, Thiago Costa, Maria da Conceição Batista, Rosalie Belian, Vaninha Vieira, Damires Souza and Carlos Eduardo Pires.

A special thanks goes to the co-authors of my papers, reports, books, book chapters, and other publications.

Finally, I would like to thank all members of my family that always supports me in my personal and professional activities. Particularly, I have a special word to two special people. Tiago, thank you for being who you are. Marco, thank you for being on my side.

Abstract

This report includes the main results in three research areas we have been working on since 1989: Geographical Databases, Data Integration and Semantic Issues in PDMS (Peer Data Management Systems).

A Geographical Database is a collection of inter-related and geo-referenced data. By definition, it is a database directed to the representation, storage and access to the information, which is spatially referenced. Traditional techniques of data modeling were not adequate for the treatment of geographical data. The difficulty consists of the fact that most of these data are validated in terms of its spatial localization, time, and the reliability of the collection. In this context, our contribution was the proposal of an object-oriented geographic data model MGeo+ and its query language LinGeo. We also have worked on spatial access methods' analysis and on the proposal of a visual query language for geographical data along with its user interface.

The data integration systems are tools that offer a uniform access to distributed and heterogeneous Web data sources. This is done by resolving the heterogeneities and giving to the disparate sources an uniform view. Users submit queries over the integrated view without having to spend a lot of time in searching and browsing the Web. We have been working on the specification and implementation of a data integration system mainly interested in the evolution of the mediation schema, query reformulation and quality issues.

Schemas and instances drawn from heterogeneous, dynamic and distributed data sources rarely contain explicit semantic descriptions which could be used to derive the meaning or purpose of schema elements (e.g. entity, attribute and relationship). Implicit semantic information needs to be extracted in order to clarify the meaning of the schema elements. To achieve this, an ontology of a given knowledge domain will provide the information regarding semantic relations among the vocabulary terms shared by the data sources. Semantic interpretation, however, regards people's understanding and it is a context-dependent task which requires a specific understanding of the shared domain knowledge. *Context* may be employed as a way to improve decision-making over heterogeneity reconciliation in data integration processes since it helps to understand the data schema semantics as well as the data content semantics. We present our proposal to a context-oriented model and a domain-independent context manager, a contextual ontology to data integration and a semantic-based approach to peers' organization in a PDMS.

Résumé

Ce rapport présente mes principaux résultats en trois axes de recherche depuis 1989 : les Bases de Données Géographiques, l'Intégration de Données et la prise en compte de la Sémantique dans les Systèmes Pair-a-pair (P2P).

Une base de données géographiques est dédiée à la représentation, au stockage et à la récupération d'informations référencées dans l'espace. Les techniques traditionnelles de modélisation n'étaient pas adéquates pour le traitement de ces types de données. La difficulté vient du fait que la plupart des données sont validées en termes de leurs localisation dans l'espace, du temps et de leur disponibilité. Dans ce contexte, notre contribution a été la proposition d'un modèle de données géographiques orienté-objet, *MGeo+,* et son langage de requête, *LinGeo*. Nous avons aussi travaillé sur l'analyse des méthodes d'accès spatiales et sur la proposition d'un langage de requêtes visuel et son interface utilisateur.

Les systèmes d'intégration de données sont des outils qui offrent un accès uniforme à des sources de données distribuées et hétérogènes. Cela est accompli en identifiant les hétérogénéités et en fournissant une vue unifiée sur les diverses sources. Les utilisateurs envoient leurs requêtes sur cette vue intégrée sans perdre du temps à naviguer sur le Web. Nous travaillons sur la spécification et l'implémentation d'un système d'intégration de données et, en particulier, sur les aspects d'évolution du schéma de médiation et de la qualité des schémas.

Les schémas et les instances des sources de données hétérogènes, dynamiques et distribuées contiennent rarement des descriptions sémantiques explicites qui puissent être utilisées pour dériver le sens des éléments du schéma (entité, attributs et associations). L'information sémantique implicite doit être extraite pour clarifier la signification des éléments du schéma. Pour permettre cela, une ontologie du domaine fournira les informations des associations sémantiques entre les termes du vocabulaire partagé par les sources. Cependant, l'information sémantique a un rapport avec la compréhension des gens et est une tâche dépendante du contexte et qui nécessite une connaissance spécifique du domaine. Le concept de contexte peut être employé pour améliorer la prise de décision afin de résoudre l'hétérogénéité sémantique des processus d'intégration de données une fois qu'il aide à la compréhension sémantique du schéma des sources et de leurs contenus. Nous présentons notre proposition d'un modèle de contextes, d'un gestionnaire de contextes indépendant du domaine, d'une ontologie d'informations contextuelles pour l'intégration de données et d'une approche pour la prise en compte des aspects sémantiques dans les systèmes pair-a-pair (P2P).

Contents

Foreword

Since my PhD studies I was interested in solutions for the emerging applications. My first interest was Multimedia Databases and the proposal of an object-oriented physical objects' manager as the kernel of a multimedia DBMS. Then, when I joined the UFPE database group I had worked in this research area in which the main contributions were a data type manager and an extension to the relational model. These researches were done within the following projects:

- Implementing a Multimedia Database based on an Extension of the Relational Model,
- Data management and Application for Multimedia Databases: The Case of Spatial Data

The growth of Geographical Information Systems, and the need to properly store and access spatial data, had motivated a large investigation to understand and propose solutions to geographical modeling and implementation problems. In this sense, the main contributions were an object-oriented approach to modelling geographic applications, a performance analysis of multidimensional access methods and a visual query language. The following projects allowed these researches:

- Modeling and Implementation of Geographical Databases
- Geotec - Geoinformatics: Methods and Techniques
- Distributed Geographical Databases and their Applications on the World Wide Web

The emergence of the Web have caused a huge interest on how to integrate heterogeneous and distributed data sources (databases and other structured or semi-structured documents). I have worked on data integration issues since then. The proposal of a mediator-based system architecture, an approach to mediation query evolution and a query reformulation process has lead to the implementation of a data integration system prototype, in the context of the project:

- Information Integration in Heterogeneous Environment: Architectures, Models and Implementations

The new distributed architecture (P2P and GRID) and the Semantic Web associated issues are our current research focus. I am currently working on evolution and quality issues of data integration systems, and also on semantic approaches applied to Peer Data Management Systems (PDMS), in these ongoing projects:

- GridVida: Unified View of Electronic Patient Records in Grid Computing
- SPEED: Semantic Peer-to-Peer Data Management System
- Evolution and Quality Management in Dynamic Data Integration Systems, International cooperation research project with partners in France, Uruguay and Brazil

In addition, I have also worked in other research areas such as Information Retrieval, Cooperative Systems, Cooperative Learning and Autonomous Databases with some other results. All the research projects had financial support of Brazilian institutions, two of them in the context of an international cooperation.

To summarize, in all these projects I supervised six concluded PhD theses, four others ongoing, and 31 MSc theses, with three ongoing. All the concluded PhD are associate professors in Brazilian universities. It is very important to say that I would not be able to accomplish the results presented in this report without the contribution of all the students (graduate and under-graduated) I supervised.

This report presents the research done since 1989. It is important to notice that the state of the art of each research area is not of nowadays but related to the respective period of time.

CHAPTER 1

Introduction

1.1 Historical Retrospective

In the late 1970's and early 1980's, when the concept of abstract data types were included in some programming languages and when personal computers emerged, the information systems' application domains was extended allowing the implementation of the so called non-conventional applications (medicine, geographical, office automation, CAD, among others). These new applications demand the use of specific data types (graphics, audio, images and others), characterizing the multimedia applications. At this time, the database community had identified limitations on the relational model, mainly to represent complex objects, and there were a large number of proposals to extend the basic relational model. In parallel, the success of object-oriented programming languages had caused the emergence of a new paradigm of database modeling and implementation, and a large number of object-oriented Database Management Systems (DBMS) prototypes appeared (O2, Orion, ObjectStore, among others).

The era of the internet in the 1990's changed the way information systems were implemented. First of all, the need to represent (HTML) and to search (search engines) the huge amount of non structured data spread in the Web. The second phase was to include more structure and semantics (XML) in the representation of Web data, and afterwards to integrate semi-structured and structured data stored on distributed, heterogeneous and autonomous data sources. Recently, Peer Data Management Systems (PDMS) came into the focus of research as a natural extension to distributed databases in the peer-to-peer (P2P) context [Herschel & Heese 2005]. PDMS are P2P applications where each peer represents an autonomous data source and exports its entire data schema or only a portion of it.

The increasing use of the Web has caused a permanent growth in the amount of data available on it. Efforts to overcome the obstacles created by this ample growth, associated with the desire of inserting some level of intelligence to the retrieval of documents disposed in the Web, have motivated the development of the new generation of the Web: the Semantic Web [Berners-Lee et al. 2001].

The rest of this chapter will present the main issues that motivated our research interests. First of all, the Multimedia Databases, especially the Geographical Databases and the problems related to spatial data modeling and implementation, which are discussed in Section 1.2. The emergence of the Web and its growth are discussed in Section 1.3. The concepts associated to the Semantic Web, and the need to integrate multi-source data (databases or other structured or semi-structured documents) are presented in Section 1.4. Finally, in Section 1.5 we present the structure of this report.

1.2 Multimedia Databases

The diversity of information systems application domains caused the need to represent complex data types (graphics, video, image, audio, text, among others) [Grosky 1994]. The programming languages in a first moment, and the DBMS later, provided functionalities to include new data types according to the application to be implemented. These multimedia data types are characterized by a huge volume and have some requirements, which have implications on their storage, manipulation, and presentation (with appropriate interfaces). A multimedia database management system must provide a suitable environment for using and managing the multimedia data [Woelk & Kim 1987]. The geographical databases are a special case of multimedia databases in the sense that they manipulate non standard data types (point, line, polygon, cells or tessellations). Spatial data are complex, voluminous and need a specific set of operators.

We were mainly interested in modeling and implementation issues of geographical databases. In this sense, we have proposed an object-oriented modeling approach to spatial data (objects and fields) [Times 1994, Pimentel 1995]. A partial implementation of this approach was done using the Postgres [Stonebraker & Rowe 1986] DBMS. The Postgres existing spatial data types were extended with new types according to the proposed model, and a query language was implemented as an extension of POSTQUEL query language [Nascimento 1995].

The management of the spatial data objects is more complex than the handling of conventional data made by traditional DBMS. This complexity issue arises as new data types are needed to describe the geometry of spatial data objects from coordinate data. To deal with this problem, DBMS were extended to include more efficient data structures and special search algorithms, known as multidimensional access methods (MAM) [Gaede & Günther 1998]. These methods have been designed for providing an optimum access path to spatial data. Nevertheless, with the great amount and diversity of proposed MAM, there was a need for comparing the efficiency of several proposals supporting both distinct data configurations and many query types. For this purpose, we have analyzed a group of multidimensional access methods (called the R-tree group) in terms of the spatial data distribution using the experimental technique of database benchmark [Ciferri 2002].

User interfaces for GIS require considerable research and technology development efforts for usability enhancement [Wessel & Haarslev 1998]. An important aspect of usability in this context is the need for a visual query language to enable users to think graphically, while building queries. In this sense, we have proposed a geographical visual query language, *GeoVisualQL* [Soares 2002], and its respective user interface, *GeoVisual Interface* [Souza 2000].

1.3 The Web

The original efforts to develop the Web as it is known today were done in the beginning of the 1990's [Berners-Lee et al. 2001]. Such efforts however, were fully rewarded, once the Web had been consolidated as the mean of information distribution with the faster growth in the worldwide history. On the other hand, the increasing use of the Web has caused a permanent growth in the amount of data available. This fast expansion made the Web a data repository as huge as confused. Its specific characteristics have generated a crescent demand for tools specialized in performing efficient management and qualified data retrieval/extraction from Web contents [Baeza-Yates & Ribeiro-Neto 1999].

The standard language used to create Web pages, the HTML (Hypertext Markup Language) [Ragget 2005] does not specify any semantics related to the resource it formats, being responsible just for the appearance of the Web document. Thus, a gap emerges between the information that is available to Web services and the one that is provided for human reading.

Search engines were the first attempt to access Web content but it mainly deals with unstructured data [Baeza-Yates & Ribeiro-Neto 1999]. These classes of software mechanisms are both guided by searches based on the meaning of the keywords provided by its users or contained in the text. These tools do not consider the semantic aspects involved in the keywords that were submitted to the search, analyzing the word just syntactically. A complementary research area is the Information Extraction [Adams 2004]. Its main objective is to extract relevant information from semi-structured documents, and present the information extracted in a user friendly format.

The lack of meaning in HTML documents makes hard to software agents processing these types of documents in an intelligent way, being necessary to develop a way that allows us to insert some "intelligence" in the Web resources. XML includes some semantics on documents through the definition of significant tags [Bray et al. 2006]. XML allows the representation of semi-structured data.

Integrated access to multiple data sources ranging from traditional databases to semi-structured data repositories was required in many applications. In the last decade, the problem of integrating data from distributed, heterogeneous and autonomous data sources has received a great deal of attention. It consists in providing a uniform view of these data sources in which the users can pose their queries. We have proposed and implemented *Integra*, a prototype of a mediator-based data integration system. *Integra* adopts the Global-as-View (GAV) approach [Halevy 2000] to define mappings between the integrated mediation schema and the data source schemas. Our main contributions were associated to the specification and implementation of two important processes: (i) the definition and evolution of mediation queries when data source schemas change [Loscio 2003], and (ii) the query reformulation process [Costa 2005]. Recently, we have been working in the definition and evaluation of schema quality criteria for the generated mediation schema [Batista 2008].

1.4 The Semantic Web

The Semantic Web can be thought as an extension of the current Web, where data gains its own meaning. The main objective of this new Web paradigm is to insert some level of knowledge into WWW resources, so that software agents can be able to intelligently process Web contents [Hendler 2001]. According to [Shadbolt et al. 2001], agents can only flourish when standards were well established and the web standards for expressing shared meaning have progressed steadily over the last years.

A multi-layer architecture has been proposed by Tim Berners Lee [Berners-Lee et al. 2001, Studer 2003] to support the Semantic Web. Amongst these layers, we would like to highlight the Ontology vocabulary. The specification of an ontology makes possible the communication between computer systems, independently of the architecture and the information domain treated, eliminating ambiguities over a domain terminology [Bézivin 1998, Cardoso 2007]. So, we can say that an ontology provides a common/shared understanding about concepts of specific knowledge domains.

Data integration fits in the Semantic Web scenario which demands for complete application interoperability. To include more semantic issues in *Integra*, we have been working on a semantic name resolution process to the mediation schema generation process [Belian 2008]. This process uses syntactic methods to match element names against terms of a domain ontology, and uses contextual information to reduce the range of similar terms obtained in the linguistic match. In this case, the contextual information provides knowledge about data source schemas and element names (terms) with this purpose. The contextual information used in this work is represented by a context ontology enabling information reusability and sharing, besides reasoning capability.

1.5 Intensive Access to Remote Data

Nowadays, we see the growth of remote heterogeneous data all over the Web and the coming of new distributed architectures to facilitate the access to these data. Peer-to-Peer (P2P) is one of these new architectures used mainly for file sharing and, more recently, for structured data sharing. Data management in P2P systems is a challenging and difficult problem considering the extreme number of peers, their autonomous nature and the potential heterogeneity of their schemas [Halevy et al. 2006]. A Peer Data Management System (PDMS) is one such application which enables users to transparently query several heterogeneous and autonomous data sources. A PDMS is considered an evolution of the traditional data integration systems [Sung et al. 2005]. As data integration systems, PDMSs realize their services over data from existing heterogeneous sources and, in order to reconcile heterogeneity, more semantic information about the data sources is needed. In this sense, Web Semantic issues such as metadata, contexts and ontologies are used to solve semantic conflicts among information from diverse heterogeneous web data sources [Kashyap & Sheth 1996].

We are currently working on the specification of SPEED (**S**emantic **PEE**r-to-Peer Data Management System) an ontology-based PDMS in which the content shared by peers (exported schema) is represented through ontologies. The system adopts a mixed network topology (DHT and super-peer) in order to exploit the strengths of both topologies. A DHT network is used to assist peers with common interests to find each other and form semantic communities. Within a community, peers are a grouped in clusters of semantically related peers. We are addressing two main issues: the formation and maintenance of clusters [Pires 2009] and the query rewriting among peers that are semantic neighbors using contextual information [Souza 2009].

Context is what underlies the ability to differentiate one situation from another and to characterize entities and events. The concept of context is been used in several domains, and mainly in context-aware ubiquitous applications. Most context-sensitive systems do not take into account requirements such as modularity, reusability or interoperability and implement context manipulation tasks in a proprietary way, so as to fulfill the particular needs of each system. Context management aims at providing solutions to separate context manipulation tasks from applications' business. We are also currently working on the specification and implementation of a domain-independent context manager based on the Context-Oriented Model [Vieira 2008]. The main idea is to provide a general framework to context-aware application developers.

1.6 Report Structure

This report will focus on three main research areas that we talked about in previous sections: Geographical Databases, Data integration and Semantic Issues in PDMS. In this sense, it includes the following content:

- Chapter 2 – Geographical Databases, where we discuss the several issues we had been working on Geographical Databases along with our main contributions in the area.
- Chapter 3 – Data Integration, which contains our proposal to a data integration system, its characteristics and processes.
- Chapter 4 – Semantic Issues in PDMS, where we discuss some semantic issues and the use of semantic-based techniques in a PDMS proposal.
- Chapter 5 – Other Activities, where we briefly present some other research areas we have been working on.
- Chapter 6 – Conclusion and Perspectives, which summarize the main issues discussed in this report and presents some perspectives for future works.

- Annexe 1, which contains a summary of the main collaborative projects we have participate
- Annexe 2, which presents a quantitative summary of the results obtained by research area.

1.6 References

[Adams 2004] Adams, K.C. The Web as Database: New Extraction Technologies and Content Management, ONLINE: *Exploring Technology & Resources for Information*, Vol.25, N.2, 2001 (also available at http://www.onlinemag.net/OL2001/adams3_01.html, last access March 2008)

[Baeza-Yates & Ribeiro-Neto 1999] Baeza-Yates, R., Ribeiro-Neto, B. Modern Information Retrieval. *ACM Press*, Nova York, 1999

[Batista 2008] Maria da Conceição Moraes Batista. Schema Data Quality on Information Integration Systems, *PhD Thesis*, Center for Informatics – UFPE, 2008 (ongoing)

[Belian 2008] Rosalie Barreto Belian. A Context-based Name Resolution Process for Schema Integration, PhD, 2008

[Berners-Lee et al. 2001] Berners-Lee, T., Hendler, J., Lassila, O. The Semantic Web*, Scientific American*, Vol.284, N.5, 2001, p. 34–43.

[Bézivin 1998] Bézivin, J. Who´s Afraid of Ontologies. In Proc. of the *Model Engineering, Methods and Tools Integration with CDIF Workshop (OOPSLA'98)*, Vancouver, Canada, 1998

[Bray et al. 2006] Bray, T., Paoli, J., Sperberg-McQueen, C.M, Maler, E., Yergeau, F. Extensible Markup Language (XML) 1.0 (Fourth Ed.), *W3C*, 2006 (available on http://www.w3.org/TR/REC-xml, last access February 2008)

[Cardoso 2007] Cardoso, J: The Semantic Web Vision: Where are We?, *IEEE Intelligent Systems*, Vol. 22, No. 5, 2007, p. 84-88.

[Ciferri 2002] Ricardo Rodrigues Ciferri. Analysing the Performance Variation of Multidimensional Access Methods in Terms of Spatial Data Distribution, *PhD Thesis*, Center for Informatics – UFPE, 2002

[Costa 2005] Thiago Alves Costa. Gerenciamento de Consultas em um Sistema de Integração de Dados, *MSc Thesis*, Center for Informatics – UFPE, 2005

[Gaede & Günther 1998] Gaede, V., Günther, O. Multidimensional Access Methods, *ACM Computing Surveys*, Vol.30, N.2, 1998, p. 170-231

[Grosky 1994] Grosky, W.I. Multimedia Information Systems, *IEEE Multimedia*, Vol.1, N.1, 1994, p. 47-59

[Halevy 2000] Halevy, A., Y.: Logic-based Techniques in Data Integration. J. Minker, editor *Logic based Artificial Intelligence*, Kluwer Publishers, 2000

[Halevy et al. 2006] Halevy A., Rajaraman A., Ordille J. Data Integration: The Teenage Years, In Proc. of the *32nd International Conference on Very Large Data Bases (VLDB*), Seoul, Korea, 2006, p. 9-16

[Hendler 2001] Hendler, J. Agents and the Semantic Web, *IEEE Intelligent Systems*, Vol.16, N.2, 2001, p. 30-37.

[Herschel & Heese 2005] Herschel, S., Heese, R. 2005. Humboldt Discoverer: A Semantic P2P index for PDMS. In Proc. of the *International Workshop Data Integration and the Semantic Web (DisWeb)*, Porto, Portugal, 2005

[Kashyap & Sheth 1996] Kashyap, V., Sheth, A.P. Semantic and Schematic Similarities Between Database Objects: A Context-Based Approach, *VLDB Journal*, Vol.5, N.4, 1996, p. 276-304

[Loscio 2003] Bernadette Farias Lóscio. Managing the Evolution of XML-based Mediation Queries, *PhD Thesis*, Center for Informatics – UFPE, 2003

[Nascimento 1995] Adriana Maria Rebouças do Nascimento. *LinGeo* - Uma Linguagem de Consulta Geográfica, *MSc Thesis*, Center for Informatics – UFPE, 1995

[Pimentel 1995] Flávio Leal Pimentel. Uma Proposta de Modelagem Conceitual Para Dados Geográficos - Modelo *MGeo+*, *MSc Thesis*, Center for Informatics – UFPE, 1995

[Pires 2009] Carlos Eduardo Santos Pires. Semantic-based Approach for Peer Clustering in a Peer Data Management System, *PhD Thesis*, Center for Informatics – UFPE, 2009 (ongoing)

[Raggett 2005] Raggett, D. Getting Started with HTML, *W3C*, 2005 (available on http://www.w3.org/MarkUp/Guide/, last access February 2008).

[Shadbolt et al. 2006] Shadbolt N., Berners-Lee T., Hall W. The Semantic Web Revisited, *IEEE Intelligent Systems*, Vol.21, N.3, 2006, p.96-101.

[Soares 2002] Valéria Gonçalves Soares. GeoVisual – A Visual Query Environment for Geographical Databases, *PhD Thesis*, Center for Informatics – UFPE, 2002

[Souza 2000] Damires Yluska de Souza Fernandes. GeoVisual Interface - Uma Interface para Consultas Visuais em Banco de Dados Geográficos, *MSc Thesis*, Center for Informatics – UFPE, 2000

[Souza 2009] Damires Yluska de Souza Fernandes. Semantic-based Query Reformulation for PDMS, *PhD Thesis*, Center for Informatics – UFPE, 2009 (ongoing)

[Stonebraker & Rowe 1986] Stonebraker, M., Rowe, L.A. The Design of POSTGRES, In Proc. of the *ACM SIGMOD International Conference on Management of Data*, Washington, USA, 1986, p. 340-355

[Studer 2003] Studer, R., Agarwal S., Volz, R. The Semantic Web Methods, Applications and Future Trends, In Proc. of the *3rd IFIP Conference on e-Commerce, e-Business, and e-Government,* Guaruja, Brasil, 2003, p.203-213

[Sung et al. 2005] Sung, L.G.A., Ahmed, N., Blanco, R., Li, H., Soliman, M.A., Hadaller, D.A. Survey of Data Management in Peer-to-Peer Systems, *School of Computer Science*, University of Waterloo, Canada, 2005

[Times 1994] Valéria Cesario Times, V. *MGeo*: An Object-oriented Model for Geographical Applications, *MSc Thesis*, Center for Informatics – UFPE, 1994

[Vieira 2008] Vaninha Vieira dos Santos. A Domain Independent Approach for Managing Contextual Elements, *PhD Thesis*, Center for Informatics – UFPE , 2008 (ongoing)

[Wessel & Haarslev 1998] Wessel, M. Haarslev, V. VISCO: Bringing Visual Spatial Querying to Reality, In Proc. of the *IEEE Symposium on Visual Languages*, Halifax, Canada, 1998, p.170-177

[Woelk & Kim 1987] Woelk, D., Kim, W., Multimedia Information Management in an Object-Oriented Database System, In Proc. of the *International Conference on Very Large Data Bases (VLDB*), Brighton, England ,1987, p. 319-329

CHAPTER 2

Geographical Databases

2.1 Introduction

A Geographic Information System (GIS) is "a powerful set of tools for collecting, storing, retrieving at will, transforming and displaying spatial data from a particular set of purposes" [Burrough & McDonnell 1998]. Geographical data represent phenomena from the real world in terms of their position in coordinate systems, their attributes and their spatial interrelations with each other.

A Geographical Database is a collection of inter-related and geo-referenced data. By definition, it is a database directed to the representation, storage and access to the information, which is spatially referenced. It must also provide the user with the capacity to: work with several and different data types, get all the relevant information in a consistent way, and access a toolkit of sophisticated analysis.

Geographical Data Modeling consists in the formulation of an adequate set of abstractions to represent the geographical reality in the database, and in the definition of data handling conditions and integrity rules.

Traditional techniques of data modeling were not adequate for the treatment of geographical data [Orenstein & Manola 1988, Ooi et al. 1989, Egenhofer 1993, Gardarin 1993]. The difficulty consists of the fact that most of these data are validated in terms of its spatial localization, time, and the reliability of the collection [Roberts et al. 1991]. According to Peuquet [Peuquet 1984], multidimensionality, large number of relationships and complex spatial definitions make the modeling of geographical data uniquely difficult. The models themselves tend to be complex and the resultant data files tend to be not very compact. The problem of a lack of uniformity to deal with geographical data was due in large part to the several differences in the commonly used storage formats. Basically, this was also due to a lack of fundamental knowledge concerning properties of spatial data and concerning the design and evaluation of spatial data models.

Geographical databases cover a very large range of applications problems. For this reason, some geographical models are directed to a specific application area [Lapalme et al. 1992, Williamson & Stucky 1992] or for the sake of simplicity they are exclusively designed to deal with operations belonging to raster or vector format. Object-oriented data models have facilities to express more suitably the knowledge domain of several geographical applications [Oosterom & Bos 1989, Dueker & Kjerne 1987, Egenhofer & Frank 1987] due to the spatial nature of their data and operators. An object model for a geographical database has spatially referenced objects as its constituents.

In what follows, we present the object-oriented geographic data model *MGeo+* and its query language *LinGeo*. We also present some results concerning spatial access methods' analysis and a proposal we had made on a visual query language for geographical data along with its user interface.

This research had financial support (four projects) from a Brazilian institution (CNPq) from 1991 to 2003.

2.2 Spatial Data Modeling

Spatial data models are grouped into three categories. The first one is *raster* (continuous space) which represents the distribution of physical entities dividing space into cells with the same format and size, each cell being associated with only one value of the physical variable being represented. The second is *vector* (discrete objects) which divides the space into spatial objects with symbols, descriptive attributes, and coordinates associated with themselves, and each spatial object corresponds to an element of the actual world. The last one is the *hybrid* approach which incorporates aspects of both formats described above.

In this sense, the rest of this section presents a hybrid conceptual data model and its associated query language.

2.2.1 The MGeo Conceptual Data Model

We have disigned an object-oriented data model for geographical applications, called *MGeo*, which allows the modeling of spatial objects (discrete elements) [Times 1994]. Afterwards, this model was extended allowing the representation of continuous elements, generating the *MGeo+* version [Pimentel 1995]. In order to formalize the main elements of this model we used the OMT methodology [Rumbaugh et al. 1991]. The structure of a geographical database is shown in Figure 2.1, where relationships among the classes of the data model are illustrated. In what follows we will briefly describe the *MGeo+* classes' hierarchy and the defined spatial operators.

Classes' Hierarchy

GEO_DB constitutes from a more generic point of view a set of geographical data. The next most important and more generic class of the hierarchy is REPRESENTATION MODEL. This class represents an aggregation of thematic layers (THEME). Each THEME is composed by GEOGRAPHIC ENTITIES and their properties and each PROPERTY can be represented through DESCRIPTIVE, SPATIAL and SYMBOLIC REPRESENTATION. This relationship allows real world objects (e.g. cities, rivers or roads) to have more than one spatial and symbolic representation. In this way, a single object which can be represented in different scales or projections will have associated with it a level of descriptive information, a geometry and a symbology in adequate proportions.

A spatial representation is an aggregation of SPATIAL ELEMENTS and each one can have a raster format (CONTINUOS ELEMENTS), represented by CELLs and TESSELATION, or a vector format (DISCRETE ELEMENTS), represented by POINT, LINE and REGION, and their sub-elements.

As an example of a *MGeo+* class definition let consider the REGION class:

REGION CLASS

The instances of this class represent the bi-dimensional spatial objects. They are defined by a sequence of lines which delimitate a region. The following operators are defined for the Region Class:

Adjacent: verify if two regions are neighbours.

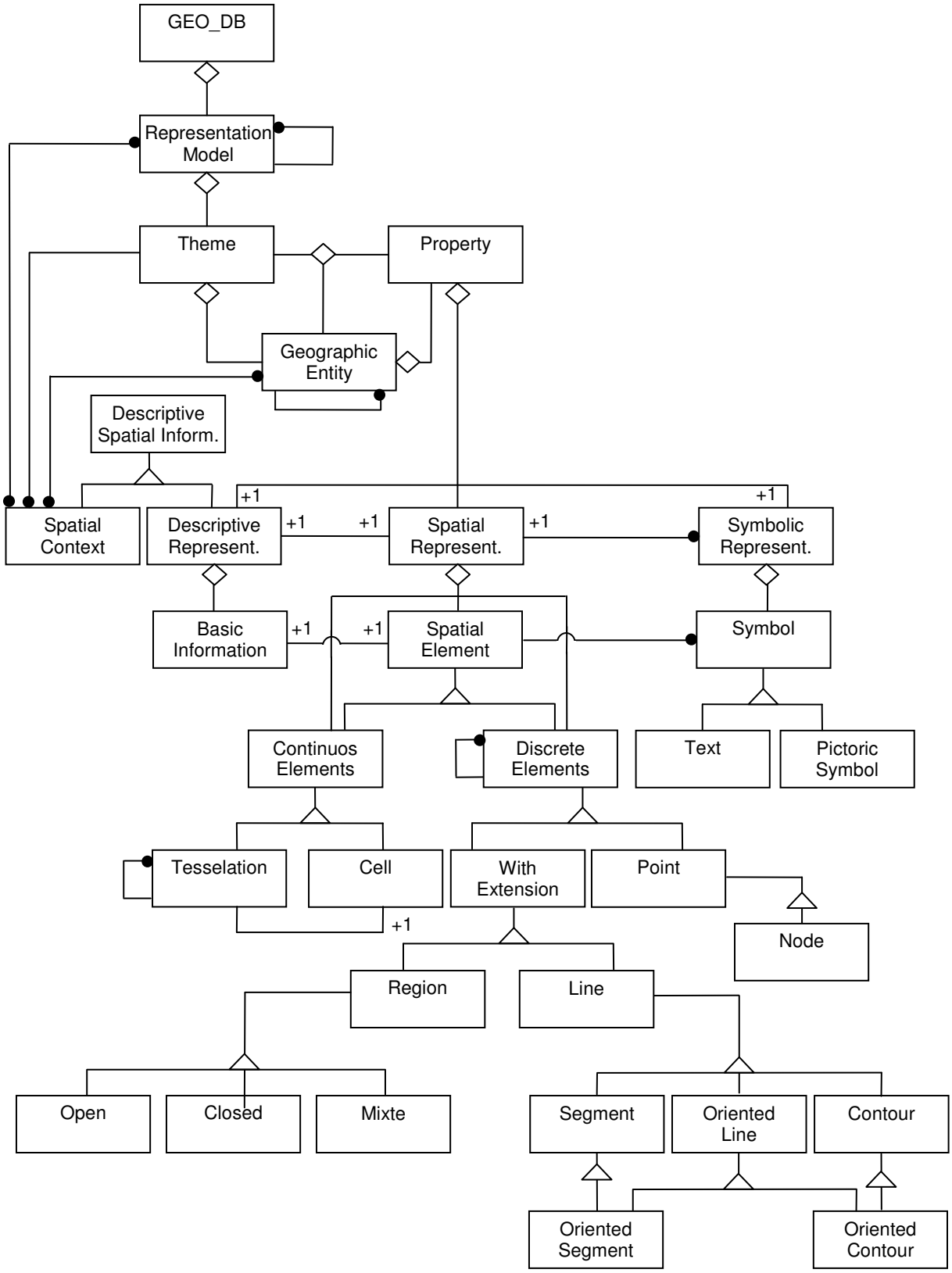


Figure 2.1 – MGeo Class Diagram

Area: calculate the area of a specific region.
Interior: return the interior of a specific region.
Component: return the topological components, which are the connected and maximum sub-regions, of a specific region.
Boundary: return the set of points which belong to the boundary of the region.

A region can have negative sub-regions. These sub-regions are internal areas of the region which are not considered by the domain of the problem being modeled at the moment. A terrain containing a lake or a region with a number of disjoint islands, which are not important to the application, can be modeled in this way. A region can also be OPEN, where its boundary is not considered; CLOSED, where its boundary is quite considered; and MIXED, where its boundary is partially considered.

Spatial Operations

Spatial operators have been defined related to spatial object classes. They are classified into three groups:

- *Set-oriented* - operators that deal with objects as set of objects: Equal, Subset, Membership, Union, Intersection, Difference, Cardinality, Complements, Elements, Windowing, Disjoint, Parts, Lines, Joint.
- *Topological* - ideal with topological properties, such as localization, adjacency and connectivity, of spatial objects: Interior, Extremes, Begin, End, Clockwise, In, Closure, Boundary, IsConnected, Connect, IsTherePath, Path, Adjacent, Component, IsBoundary
- *Metric* - deal with metric properties of space in which objects are defined. This group also includes notions of the Euclidian Metric: Distance, Length, Parameters, Direction, Area

The complete specification of all classes of *MGeo+* data model, including operations, can be found in [Times 1994, Pimentel 1995].

2.2.2 LinGeo – A Geographical Query Language

A spatial query language must allow the manipulation and access to spatial and non spatial data. In our approach we only considered vector data: point, line and polygon. Several types of spatial query languages were proposed: based on the graphical representation as MapQuery [Frank 1982], GEOQL [Sacks-Davis et al. 1987, Ooi et al. 1989], Pictorial SQL [Roussopoulos & Leifker 1985] and Spatial SQL [Egenhofer 1994], based on QBE (Query-by-Example): QPE [Chang & Fu 1980] and PICQUERY [Joseph & Cardenas 1988]; and based on Quel: POSTQUEL [Stonebraker & Rowe 1986] and Geo-QUEL [Berman & Stonebraker 1977].

We have proposed *LinGeo* [Nascimento 1995] whose objective was to define and manipulate spatial objects as stated by *MGeo+*. One of its main characteristics was the extensibility, i.e., to allow de definition of new data types, operators and functions. *LinGeo* was implemented as an extension of the spatial data types proposed by PostGres [Stonebraker & Rowe 1986] (point, lseg, line and polygon) with new data types: cont (contour), cont_or (oriented contour), line_or (oriented line) and lseg_or (oriented line segment). Some of the *MGeo+* spatial operators were also implemented extending POSTQUEL.

LinGeo was created to facilitate the management of geographical data allowing the creation and deletion of spatial classes; the update, inclusion, exclusion and querying of geographical data; the manipulation of spatial and non spatial data in a similar way; and the offer of geographical operators.

2.2.3 Summary

The proposed *MGeo+* modeling approach was the base of several works that have been developed in our database group, and also in cooperation projects with other institutions. It was completely formalized using MooZ (Modular Object-Oriented Z) [Meira & Cavalcanti 1992]. MooZ is a language that enriches the formal language Z [Bowen 1992] with object-oriented concepts and whose semantics is based on set theory and first-order logic. The *MGeo+* and its *LinGeo* language were part of three MSc theses [Times 1994, Times & Salgado 1994, Pimentel 1995, Nascimento 1995, Nascimento & Salgado 1995]

The master student that proposed *MGeo* first version has continued her PhD studies in Leeds, England, and is currently the leader of the geographical database researches at UFPE. We are both supervising a PhD thesis that proposes a query language to integrate geographical and analytical (OLAP) operators [Silva 2008].

2.3 Performance Analysis of Spatial Access Methods

Spatial Database Management Systems (SDBMS) extend the typical functionalities used for the management of alphanumeric data in secondary memory by integrating and supporting spatial data types (e.g. points, lines, polygons and three-dimensional solids), in their data model, query language and physical implementation [Güting 1994]. However, the query processing performance of these systems has strongly been influenced and improved through the use of more efficient data structures and special search algorithms, known as multidimensional access methods (MAM) [Gaede & Günther 1998]. These methods have been designed for providing an optimum access path to spatial data which is based on a well-defined set of predicates. In this sense, the indexed space is organized in such a way that allows the retrieval of spatial data objects that are located close to each other. For example, the execution of queries about spatial objects contained in a certain area just requests the access to objects located close to this area, as opposed to the analysis of the whole set of objects stored in the disk.

Several factors influence MAM performance, which can be associated with the data, the workload, the tuning parameters of the data structure and the buffer-pool management. These factors are called *decisive performance factors*. Among these, the spatial data distribution is one of the factors that strongly affect MAM performance as it directly affects the way most access methods partition space. In other words, it influences the formation of areas that divide space and, therefore, it affects the data clustering task of an access method and might degenerate its internal organization.

This work investigated the performance variation of a certain group of multidimensional access methods (called the R-tree group) in terms of the spatial data distribution. By evaluating the efficiency of these access methods concerning 21 different datasets, we verify how the spatial data distribution affects the costs of both inserting new entries in the data structure and making spatial query selections. In order to appropriately run the performance tests, we devised a methodology that allows the generation of a set of data distributions with various characteristics. This set enabled us to analyze the spatial data distribution factor from distinct perspectives.

The next subsections present the main characteristics of the workbench, as well as some aspects of the R-tree access methods that were considered in our work, and the performance results obtained from our study.

2.3.1 Experimental Setup

The performance tests were executed on Sun SparcStation 4 workstations under the operating system UNIX, having 320 Mbytes of RAM memory and 4.5 Gbytes of disk. In order to be able to compare our work with the performance results reported by other researchers, especially by Cox [Cox 1991] and Carneiro [Carneiro & Magalhães 1998], our performance results were measured according to the number of disk accesses (i.e. number of disk pages accessed during the execution of insertion operations and spatial queries). This measure particularly aimed to avoid the collateral effects caused by the concurrent execution of several programs and applications of a multi-user and multi-programming environment which is an aspect inherent to our testing environment. This also allowed us to determine the costs of data I/O from/to disk, which is in turn the main part of the costs involved in the execution of most of the spatial selection queries.

2.3.2 Multidimensional Access Methods Addressed

The investigated multidimensional access methods, called the R-tree group, consist of the following methods: R-tree with quadratic-cost split algorithm [Guttman 1984], the R^+-tree [Sellis et al. 1987] and three variants of the R^*-tree. The first variant of the R^*-tree is an extension of the original work, as reported in Beckmann. [Beckmann et al. 1990]. This extension uses a routine for reinsertion of entries during the operation of data insertion, when the first overflow treatment occurs in a certain hierarchical level of the tree (forced reinsert). According to the original work, the process of reinsertion of entries employs the *close reinsert* technique. This technique asserts that among p chosen entries (i.e. entries having MBB with the most distant centers from the center of the MBB of the node), the reinsertion is made from the less distant entry to the most distant one. The second variant, called by Carneiro [Carneiro & Magalhães 1998] R^*-treeFR, uses the reinsertion of entries according to the *far reinsert* technique (i.e. from the most distant entry to the less distant one). Finally, the third variant of the R^*-tree, called R^*-treeSR, does not use reinsertion of entries in the insertion routine.

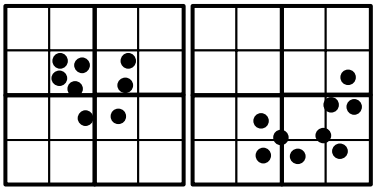
A detailed description about the structure of these multidimensional access methods can be obtained from the references listed above and from the survey conducted by Gaede and Günther [Gaede & Günther 1998]. However, it is important to discuss the possible values that the tuning parameters of the MAM data structures can take. These parameters affect the efficiency of the insertion operations and spatial queries, as acknowledged in [Kriegel et al. 1989], and they are: (i) m (i.e. minimum allowed number of entries per node or disk page) and (ii) p (number of entries to be reinserted in the overflow treatment). The goals of using the parameter m are to guarantee a minimum node occupation and consequently, to minimize storage costs. For this parameter, we used 40% of the maximum number of entries per node, for almost all access methods, except for the R^+-tree. For the R^+-tree, the value of m is equal to one. For the parameter p, which is specifically valid for the variants of the R^*-tree, we used 30% of the maximum number of entries per node [Beckmann et al. 1990].

2.3.3 Further Characteristics of the Performance Tests

The buffer-pool consisted of a set of 20 disk pages of 4 Kbytes stored in main memory and aims to minimize the number of disk accesses. The buffer-pool management was made according to the LRU policy (*least recently used*) for disk page replacement. In our workbench, the access methods or any other programs did not share the buffer-pool. For the R-tree and the three variants of the R^*-tree access methods, the 4K size for a disk page allowed a maximum number of 102 entries per node (parameter M), while for the R^+-tree, this number was reduced to 56 entries per

node. For the R^+-tree, the maximum number of entries is smaller than the others due to the storage of two rectangles for its entries: these are the MBB and the MaxRect.

The data used in the performance tests were composed by rectangles occupying 0.1% of the total area of the extent in size and linear in shape. This shape varied uniformly according to the following parameters: (i) linear-x, which specifies that the extension in dimension *x* is twice larger than the extension in dimension *y*, and (ii) linear-y, which determines that the extension in dimension *y* is twice larger than the extension in the other dimension *x*. Each data file was composed by 10,000 rectangles that had artificially been generated according to the characteristics mentioned above. Also, they were located inside of an extent limited to the unitcube $[0,1)^2$, and their spatial distribution was obtained by using the proposed methodology. This methodology allows the generation of a set of 126 types of data distribution which results from the combination of location subclasses (collective location of the data) and organization classes (data organization). The points related to the distribution are generated according to the chosen organization class that falls within the limits specified by the location subclass, as shown in Figure 2.2. In particular, these distribution types have different characteristics which enable analyzing the consequences of spatial data distribution under distinct perspectives. These perspectives range from a weak to a strong influence on the performance of MAM. However, only a subset of the possible types of data distribution that can be generated using this methodology was initially considered in our investigation. These are shown in Table 1.



Subclass 15 Subclass 14

Figure 2.2 - Examples of the Fourth Organization Class

In the total, 21 types of data distribution related to the subclasses of location 2, 3, 8, 12, 14, 16 and 21, and to the first, second and fourth organization classes were considered in our experiments. In choosing the location subclasses, we selected a representative for at least a subclass of each location class, as well as we reproduced some distribution types that had been analyzed in related works.

Table 1 - Data Files and their respective Spatial Data Distributions

Data Files	Types of Data Distribution		Data Files	Types of Data Distribution	
	Location Subclass	Organization Class		Location Subclass	Organization Class
A01	02	1	A12	12	4
A02	02	2	A13	14	1
A03	02	4	A14	14	2
A04	03	1	A15	14	4
A05	03	2	A16	16	1
A06	03	4	A17	16	2
A07	08	1	A18	16	4
A08	08	2	A19	21	1
A09	08	4	A20	21	2
A10	12	1	A21	21	4
A11	12	2			

After generating the data files, for each access method, the performance tests were made by building the indexing structure associated with each one of these files. To achieve this goal, successive insertion operations of new entries were applied to the data structure from an empty tree. Following this, we ran some types of spatial selection queries, such as *point query*, *intersection range query* and *enclosure range query*[1]. These three query types were independently executed, for each multidimensional access method and dataset investigated in our analysis. To control the execution of queries, a strategy based on four types of query files (i.e. Q_1, Q_2, Q_3 and Q_4) was employed. The files of types Q_1 and Q_2 were used for executing *point queries* and they stored point data related to 1,000 queries. While for Q_1, these point data were generated by applying an independent uniform distribution to each dimension, for Q_2, the point data followed a data-correlated distribution, making both queries and data to have very similar spatial distributions.

The files of types Q_3 and Q_4 were used for representing data relative to query windows that were employed in the *range queries*. For these files, 1,000 rectangles that are 0.01% of the total area of the extent in size and quadratic in shape were defined. The spatial distribution of the query windows was uniform and data-correlated for Q_3 and Q_4, respectively. In the total, 735 performance results were gathered. In addition, while 105 of this total of outcomes were obtained from the building phase of the indexing structure, another 630 were obtained from the execution phase of spatial selection queries.

2.3.4 Performance Results

In this section, the main issues concerning the results obtained from our performance experiments are presented and discussed. The first variant of the R*-tree is represented as R*-treeCR, and the performance results have been normalized in terms of the efficiency of this access method. The results specifically indicate an average of the results obtained from the 21 different data files shown in Table 1. In particular, the calculation of this average was derived from the analysis of the relative performance of the access methods (i.e. outcomes normalized in terms of the R*-treeCR) for each one of the data distributions. For each access method, the presented tables show the average together with the columns *B* and *W*, which respectively indicate the number of data files that are responsible for the best and worst performance result.

Table 2 presents the performance results obtained from the building phase of the indexing structure, for each multidimensional access method. These results indicated that the efficiency of the R+-tree was significantly degraded due to the use of the *clipping* technique. When inserting a rectangle, this technique may have requested the allocation of more than an entry in the leaf nodes. Consequently, the R+-tree achieved the worst performance results, requesting approximately up to 2,500% more of disk accesses.

In general, the other access methods presented very similar performance outcomes for most of the data files, but the R-tree yielded a slightly higher performance result. Cox [Cox 1991] and Carneiro's [Carneiro & Magalhães 1998] results indicated that the routine for reinsertion of entries degrades the efficiency of the R*-tree in the insertion of rectangles. According to their experimental outcomes, while it is clear that both access methods R-tree and R*-treeSR performed well, the variants of the R*-tree that used the reinsertion routine presented poor performance results. However, our results showed that the reinsertion routine does not affect the

[1] given a k-dimensional iso-oriented rectangle $\boldsymbol{R}$ □ E^d, find all data objects $\boldsymbol{o}$ enclosing $\boldsymbol{R}$. Formally: $\boldsymbol{ERQ}(\boldsymbol{R}, \boldsymbol{dataset}) = \{ \boldsymbol{o} \mid \boldsymbol{o} \square \boldsymbol{dataset} \square \boldsymbol{o.G} \square \boldsymbol{R} = \boldsymbol{R} \} = \{ \boldsymbol{o} \mid \boldsymbol{o} \square \boldsymbol{dataset} \square \boldsymbol{o.G} \square \boldsymbol{R} \}$

efficiency of the R*-tree in the process of inserting rectangles. Similar to the conclusions reported by Beckmann [Beckmann et al. 1990], our experimental results related to the R*-tree performance indicated that the use of the reinsertion routine does not affect its efficiency as we have just stated above. Differently from Beckmann., in our performance tests, conversely, both methods R*-treeCR and R*-treeFR obtained the best performance result in just a single situation. The use of different types of data distribution allowed us to examine both: (i) a performance variation of up to 3,500% in the realm of access methods performance analysis, and (ii) some differences among the relative performance values of these methods.

Table 2 - Performance results obtained from the building phase of indexing structure

MAM	%	*B*	*W*
R-tree	**99.23**	18	0
R+-tree	802.13	0	21
R*-treeCR	100.00	1	0
R*-treeFR	100.31	1	0
R*-treeSR	101.71	1	0

Table 3 gives the performance results for the multidimensional access methods that were obtained from executing queries of type *point query*. These results indicated that the efficiency of both methods R+-tree and R-tree was reduced for most of the data distributions, and that while the first organization class exerts a great influence on the R-tree performance, the R+-tree is strongly affected by the classes of organization two and four.

The outcomes for the other access methods are rather more mixed than for the R-tree and the R+-tree. These results varied according to the data file used and the chosen type of query file (i.e. uniform or data-correlated). On average, while the method R*-treeSR performed well when the query file Q_1 was used, the R*-treeCR provided better performance results for the query file Q_2. In general, the difference among the performance results obtained for the variants of the R*-tree was very small. Again, we verified that the use of different types of data distribution leads to a great variation on the absolute performance values of the MAM (for example, for the R*-treeCR, a performance variation of more than 1,000% was identified in our experiments). In particular, for the R+-tree, our results differ from the good performance outcomes derived from the execution of *point queries* that are reported in Carneiro [Carneiro & Magalhães 1998]. The difference found in these two research efforts is due to the use of: (i) various types of data distribution, and (ii) a distinct classification of spatial object sizes, which is an important and decisive performance factor.

Table 3 - Execution Phase of Queries (*point queries*)

MAM	Q_1 (uniform)			Q_2 (data-correlated)		
	%	*B*	*W*	%	*B*	*W*
R-tree	238.62	3	10	377.05	2	12
R+-tree	271.81	0	11	1,053.88	1	9
R*-treeCR	100.00	4	0	**100.00**	7	0
R*-treeFR	98.08	7	0	102.97	8	0
R*-treeSR	**96.41**	12	0	100.69	13	0

For *intersection range queries*, while the R*-treeFR provided the best average performance results with the use of the query file Q_3 (see Table 4), the R*-treeCR presented the best outcomes with query files of type Q_4. Regarding the R*-treeSR, this access method showed to be the most

efficient approach for most of the data files used and for both types of query files Q_3 and Q_4. However, on average, the R*-treeSR showed a slightly lower performance result of approximately between 6% and 18% less. Particularly, we verified that for the query file Q_3, the efficiency of the R*-treeSR is strongly affected by the fourth organization class, causing 50% more of disk accesses.

The collective location of the data also affected MAM performance investigated in our work. For instance, the access methods R-tree and R*-treeFR obtained a performance variation of 132% and 166%, respectively, for the location subclasses two and twelve (data files A01 and A10). The R-tree performed well for the following data files: A06, A09 and A18. However, for the other files, this method did not show good performance results. Finally, the access method R+-tree performed badly for all cases examined here. In addition, this approach presented the worst performance outcomes for a large number of data files with both types of query files Q_3 and Q_4.

Table 4 - Execution Phase of Queries (*intersection range queries*)

MAM	Q_3 (uniform)			Q_4 (data-correlated)		
	%	*B*	*W*	*%*	*B*	*W*
R-tree	235.09	3	7	394.57	3	11
R+-tree	357.80	0	14	2,168.73	0	10
R*-treeCR	100.00	4	0	**100.00**	11	0
R*-treeFR	**95.92**	8	0	115.05	6	0
R*-treeSR	96.28	13	0	118.42	9	0

For *enclosure range queries*, we verified that all variants of the R*-tree provided performance results that are similar to each other. These are shown in Table 5. In particular, the performance difference among the R+-tree and the variants of the R*-tree decreased reasonably for *enclosure range queries*. For the R-tree, this was also found to be the case, although less intensively.

Table 5 - Execution Phase of Queries (*enclosure range queries*)

MAM	Q_3 (uniform)			Q_4 (data-correlated)		
	%	*B*	*W*	*%*	*B*	*W*
R-tree	201.78	0	14	213.41	1	14
R+-tree	193.97	0	7	461.84	0	7
R*-treeCR	100.00	4	0	**100.00**	4	0
R*-treeFR	**97.25**	12	0	114.46	11	0
R*-treeSR	99.34	9	0	128.14	7	0

2.3.5 Summary

The performance results collected from our experiments showed that the spatial data distribution exerts a great influence on both absolute and relative performance outcomes of the MAM. In particular, new performance relationships were identified in our study, some of which contradicted the results and conclusions obtained from previous work, such as Cox [Cox 1991] and Carneiro [Carneiro & Magalhães 1998]. Finding these new relationships was possible particularly due to the use of a methodology for generating sets of data distribution. This methodology, which is based on the spatial data placement, is quite broad and allows the generation of data distributions with different characteristics. This generation allowed us to analyze the decisive performance factor related to the spatial data distribution under distinct

perspectives, and represents the initial contributions for the definition of a standard database benchmark for evaluating the performance of MAM. This work was done as part of Ricardo Ciferri's PhD thesis [Ciferri 2002, Ciferri et al. 2003, Ciferri & Salgado 2001a, Ciferri & Salgado 2001b, Ciferri & Salgado 2000a, Ciferri & Salgado 2000b, Ciferri et al. 2000].

2.4 Geographic Visual Queries

Most current GIS are able to generate graphical representations of query results, but only a few systems support graphical queries [Danko 1997]. It is difficult for the typical GIS users to formulate queries using complex spatial operators and related entities in a textual form. This difficulty can be reduced by the use of Visual Query Languages (VQL) [Calcinelli & Mainguenaud 1994] and it is the motivation behind using visual queries in GIS.

Generally GIS interfaces have several differences from each other and they may offer too complex mechanisms to define queries. The proposed framework used a Geospatial Metadata Standard [FGDC 1999] and Spatial Query Language [OGC 1999] concepts to define a Geographic Data Visual Standard (*GeoVisual Standard*). The motivation behind using this GeoVisual Standard was designing user interfaces that are rather easier to use by improving data visualization and integration. This standard was defined based on visual elements and spatial operators and was used to define the Geographic Visual Query Language (*GeoVisualQL*). The proposed GeoVisualQL was, in turn, syntactically and semantically defined to support the main Spatial Query Language structures.

Users will not need to learn the internal structures and the query language for each GIS being used. Instead, they will use the same visual language, based on metadata standard, to query several GIS. The proposed framework supports not only visual queries by the direct manipulation of visual elements, but also enables users to navigate through the metadata information at the interface, before building the queries. In what follows we will present the proposed *GeoVisualQL* [Soares 2002] and *GeoVisual Interface* [Souza 2000].

*2.4.1 Geographic Visual Query Language (*GeoVisualQL*)*

Visual Query Systems (VQS) are database query systems that use a visual representation to depict the domain of interest and to express related requests. The goal of a typical VQS user is to retrieve the desired data by developing two main activities: understanding the target domain and formulating the query. The query formulation can be done by four strategies that identify the VQS type: by schema navigation; by sub-queries, by matching and by range selection [Catarci et al. 1997].

The main objective of the proposed framework is to be as general as possible, so that the same interface can be used in several different GIS. For this purpose, users should be allowed to use exactly the same query format in any GIS. The Geographic Visual Query Language (*GeoVisualQL*) is defined using visual elements and spatial operators with syntactic rules based on the SQL Specification of the Open GIS Consortium [OGC 1999]. This allows all possible SQL structures to be constructed in the visual language as well. Spatial operators are also graphically represented in our system to allow users to compose queries in a completely visual form.

The GeoVisual Framework is composed basically of four modules: GIS Components, Metadata Model, Query Manager, and Graphic Interface, as shown in Figure 2.3. The Query Manager Module is composed by the Geographic Visual Query Language – *GeoVisualQL*, and the Query Translator that converts the visual query to a textual query based on the SQL specification of the Open GIS.

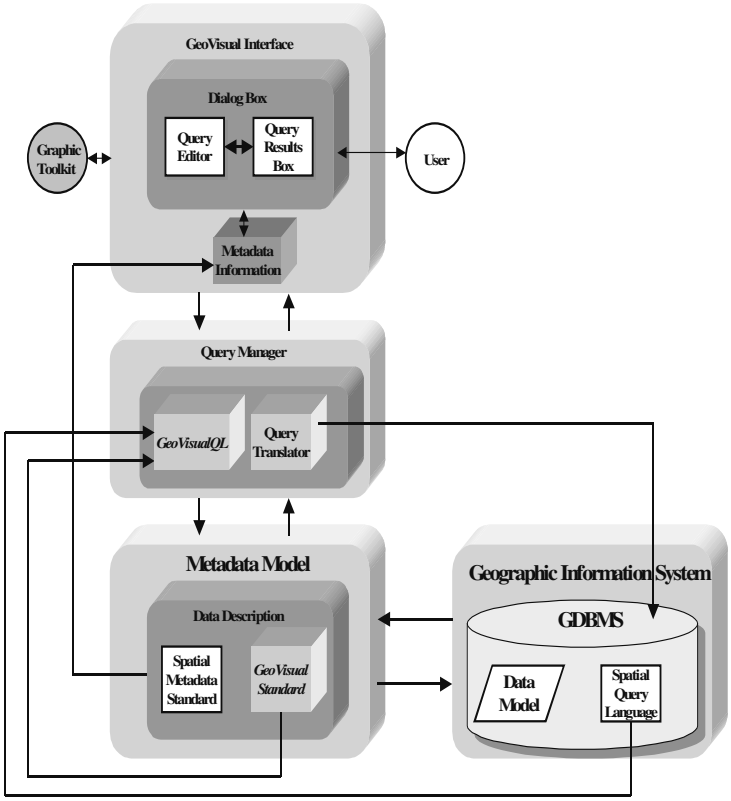


Figure 2.3 - GeoVisual Framework

In order to formalize *GeoVisualQL*, some definitions have been specified:

Definition 2.1: The *GeoVisualQL* is a triple GQL = (O, R ,Σ), where O is a set of graphic primitives that represent geographic objects, R is the set of spatial relationships allowed and Σ is the set of restrictions. *Pictorial Objects* are geographic pictoric entities identified and with attributes. They can participate of spatial relationships that define relationships between objects and sub-objects.

Definition 2.2: A pictorial Object in GQL is a triple gql = (id, o, l) where id ε ID is the object identity derived from the set of identifiers ID, o ε O is its type and l (eventually null) is the values list of its attributes.

Definition 2.3: A pictorial relationship in GQL is a 6-tuple rp = (r, po1, po2, s1, s2, s3), where r ε R is the spatial relationship type, po1 and po2 are both object identifiers or variable names and each sj (eventually null) is a set of object identifiers or variable names.

Pictorial Objects

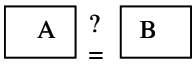
The pictorial objects of the *GeoVisualQL* and the Geographic Data Visual Standard, GeoVisual Standard, are intrinsically related to the chosen spatial data model. The primitive graphic elements (i.e. the pictorial objects of *GeoVisualQL*) are based on the SQL Specification of the Open GIS Consortium Geometry Model [OGC 1999]: point, curve, line-string, line, line-ring, polygon, multi-curve, multi-polygon, multiline-string and multipoint.

Spatial Relationships

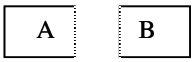
The spatial relationships and spatial operators used in *GeoVisualQL* with their respective restrictions are presented as follows. Consider the term *any* to represent any pictorial object defined above.

Definition 2.4: The following spatial relationships, and the type returned from each function compose the *GeoVisualQL* Language:

Equals *(any, any) : integer*



Disjoint *(any, any) : integer*

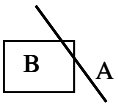


Touches *(any, any) : integer*

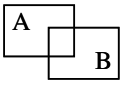
Within *(any, any) : integer*

Overlaps *(any, any) : integer*

Crosses *(any, any) : integer*

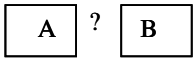


Intersects *(any, any) : integer*



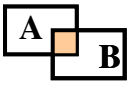
Contains *(any, any) : integer*

Relates *(any, any, string) : integer*

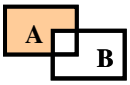


Definition 2.5: The following spatial operators also compose the *GeoVisualQL* Language. The resulting type is a pictorial object.

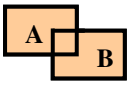
Intersection *(any, any) : any*



Difference *(any, any) : any*



Union *(any, any) : any*



Auxiliary operators had to be defined in *GeoVisualQL* to enhance the visual queries edition. The approach used in these editions, always consists of assigning pictorial objects to a given spatial operator. To allow complex queries having more than one spatial operator and *n* pictorial objects to be executed we included logic operators to be used in these queries: AND (∧), OR (∨) and NOT (¬). To group elements in the query construction we defined an auxiliary edition operator to be used if necessary, the *Group* edition operator. It is used to select which geometry entities are related with each chosen visual operator. Figure 2.4 illustrates an example of use of the *Group* operator.

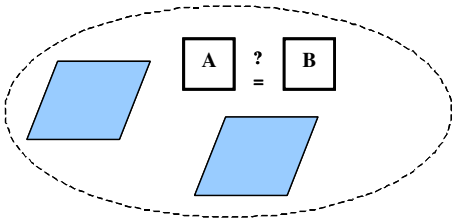


Figure 2.4 – Example of Use of the Group Operator

2.4.2 GeoVisual Interface

A spatial query formulation is a very important activity in the process of geographic data exploration. There are several ways users can build a spatial query including: item (operators, values) selection, composition of elements visualized on the screen or composition of objects presented in the database schema.

Considerable attention had been given to the relationship between spatial query language capabilities and users' perception by GIS researchers [Meyer 1992, Calcinelli & Mainguenaud 1994, Oliveira & Medeiros 1995]. It has been suggested that the interface design for querying geographic databases should take into account the following requirements [Shneiderman 1998, Oliveira & Medeiros 1995]: at which level is taken the user's initial interaction with GIS; if it is possible to think and formulate queries in a spatial way; what kind of spatial operators are available; if the database schema can be manipulated at the interface level; which help and error prevention mechanisms are provided and, finally, what kind of ergonomic approaches can be used in the interface model design.

We had proposed a visual query interface, the *GeoVisual Interface*, designed as an effort to put in an easy-to-use visual form the task of formulating queries on geographic databases [Souza 2000]. This interface is the upper level module of the proposed GeoVisual Framework [Soares 2002] and provides results derived from a multi-disciplinary study consisting of the following research subjects: query interface requirements, platform independence, spatial operators, GIS and ergonomic and human-computer interaction factors.

The designed interface has the following goals: (i) users can be novices or experts, but our main purpose is to design an easy-to-use interface for the less experienced users, (ii) the interface

should be capable of providing query easiness and geographic data exploration, and (iii) the environment where it will be developed should be platform independent and be able to be extended for use in the WEB.

In the *GeoVisual Interface* users are provided with two ways of representing their spatial information: (i) a diagrammatic model of the database schema, and (ii) the iconic representation of the spatial elements (geographic entities and spatial operators). The first option is mostly used by database experts, but may help novice users in understanding the application context, if they have any database modelling skills. The second option is really enjoyed by novice users, because it provides more familiar and intuitive elements which makes the interface more attractive and friendly. *GeoVisual Interface* adopts a hybrid strategy for formulating queries. It allows schema navigation and querying, selection of both geographic entities and spatial operators for the query construction and reuse of formulated queries.

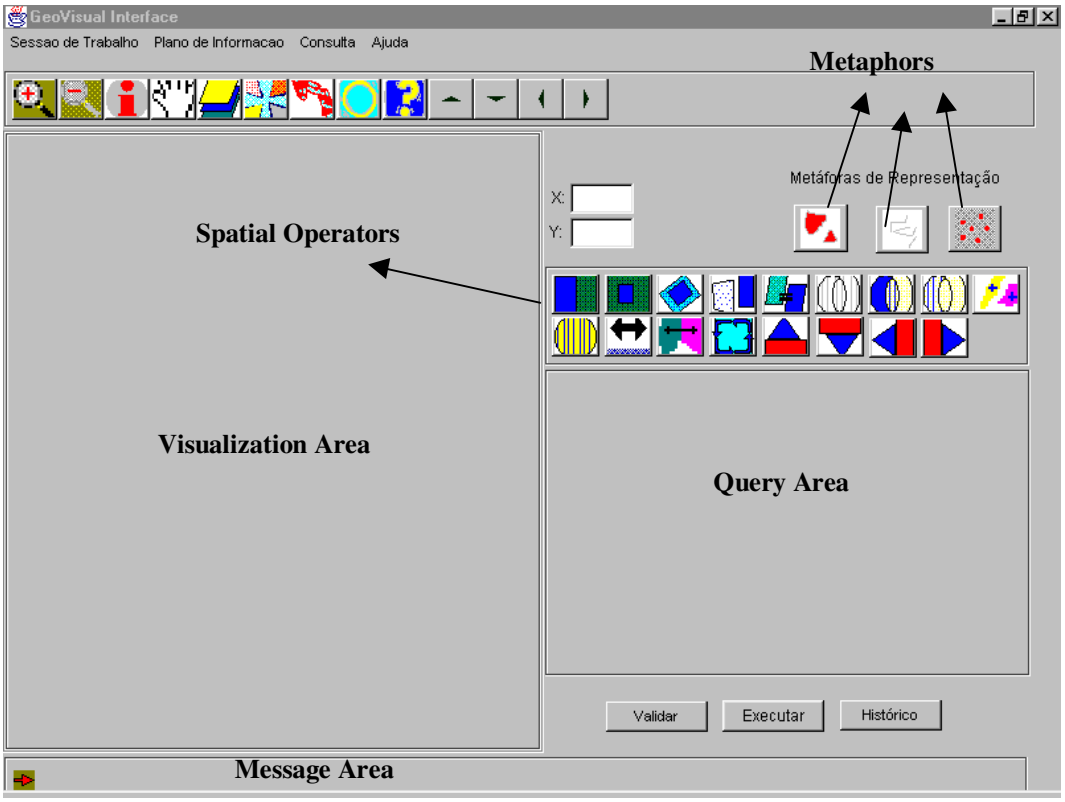


Figure 2.5 – *GeoVisual Interface* Basic Screen

Figure 2.5 presents the main window of the *GeoVisual Interface*. It contains three basic areas (i.e. the visualization, query and message areas), the pictorial representations of spatial operators and metaphors, and the principal functions used to explore geographic data, represented as icons too. More specifically, each area has a special purpose: (i) *visualization area* is used for exploration of geographic data objects or selection of geographic data for query formulation, (ii) *query area* is used for query composition, and (iii) *message area* is used to show tips and error messages. Another important area (that is optionally visualized) is called *conceptual schema area* which allows users to have access to the application database schema. Metaphors are used to represent entity categories and provide access to the list of point, line and polygon instances found in the active geographic application. Then, users can select a geographic object through this list for query composition.

From the functional point of view, our *GeoVisual Interface* provides:

a) Exploration of geographic objects.
b) Access to the application conceptual schema, using the diagrammatic visual formalism *MGeo+*/UML[Pimentel 1995, Rumbaugh et al. 1999].
c) Spatial query formulation.
d) Presentation of tips and constructive messages during the execution of each user task.
e) *GeoVisual Interface* layout designed according to VQS approach and to ergonomic design principles.

2.4.3 Summary

Based on the FGDC (FGDC Standards Reference Model, 1999) description of metadata, we proposed a Geographic Data Visual Standard, *GeoVisual Standard* and a new Geographic Visual Query Language, *GeoVisualQL*. The grammar of *GeoVisualQL* was completely formalized using the Picture Layout Grammar – PLG [Golin & Reiss 1990, Golin 1991]. This work was done as part of Valéria Soares's PhD thesis [Soares 2002, Soares & Salgado 2002, Soares & Salgado 2000, Soares & Salgado 1999]

Our proposed graphic interface has been designed to allow spatial query easiness. It has been implemented using the Java language to be completely platform independent. One important problem in GIS interfaces is how to support user queries without requiring specialized knowledge from the underlying database, or specifically from the query language. *GeoVisual Interface* provided a simple, intuitive and didactic environment to formulate visual queries in Geographic Information Systems. It is actually a user dialog module that uses *GeoVisualQL* as a background formalism. This work was done as part of Damires Souza's MSc thesis [Souza 2000, Souza & Salgado 2000a, Souza & Salgado 2000b].

2.5 Main Results

The main results related to the researches done about geographical database systems presented in this report can be summarised as:

- a database modeling approach *MGeo* [Times 1994] and its extension *MGeo+* [Pimentel 1995] with the corresponding query language *LinGeo* [Nascimento 1995] implemented as an extension of POSTQUEL;
- a spatial access methods' analysis with results about the performance variation of a certain group of multidimensional access methods (called the R-tree group) in terms of the spatial data distribution [Ciferri 2002];
- a GeoVisual Query Language, *GeoVisualQL* [Soares 2002] defined using visual elements and spatial operators with syntactic rules based on the SQL Specification of the Open GIS Consortium, and its user interface, *GeoVisual Interface*[Souza 2000].

In what follows, we present the related PhD and MSc theses concluded and the main related papers published in international and Brazilian conferences.

2.5.1 PhD and MSc Theses

We have the opportunity to work with several students in this research area resulting in three PhD and eight MSc concluded theses. They are listed in chronological order as follows.

[Times 1994] Valéria Cesario Times, V. *MGeo*: An Object-oriented Model for Geographical Applications, MSc, 1994

[Nascimento 1995] Adriana Maria Rebouças do Nascimento. *LinGeo* - Uma Linguagem de Consulta Geográfica, MSc, 1995

[Pimentel 1995] Flávio Leal Pimentel. Uma Proposta de Modelagem Conceitual Para Dados Geográficos - Modelo *MGeo+*, MSc, 1995

[Batista 1997] Daniela Coelho Freire Batista. Incorporando Características Espaciais a um SGBDOO, MSc, 1997

[Perez 2000] Celso Roberto Perez. Integration and Interoperability of GIS through Hypermedia Open Systems, PhD, 2000

[Souza 2000] Damires Yluska de Souza Fernandes. GeoVisual Interface - Uma Interface para Consultas Visuais em Banco de Dados Geográficos, MSc, 2000

[Beltrão 2001] Vicente de Paula C. R. Beltrão. Guia Metrópole: Um Sistema de Roteamento de Veículos utilizando um SIGWeb, MSc, 2001

[Ciferri 2002] Ricardo Rodrigues Ciferri. Analysing the Performance Variation of Multidimensional Access Methods in Terms of Spatial Data Distribution, PhD, 2002

[Soares 2002] Valéria Gonçalves Soares. GeoVisual – A Visual Query Environment for Geographical Databases, PhD, 2002

[Lopes 2003] Weyler Nunes Martins Lopes. Definição de Operadores de Distância Qualitativa para Objetos Geográficos Estendidos, 2003 (co-advisor)

[Melo 2005] Jonas Bezerra de Melo Júnior. Interoperabilidade de SIG Através de Serviços Web, 2005 (co-advisor)

[Silva 2008] Joel da Silva. A Geographical and Multidimensional Query Language, 2004-2008 (co-advisor, ongoing)

2.5.2 Publications

On this topic we have published twenty-five conference papers and one submitted paper to *Information Systems* journal. Among those on international forums, we would like to highlight the conferences *IFIP Conference On Visual Database Systems (VDB6)*, *ACM GIS (Symposium on Advances in Geographic Information Systems)* and *ACM SAC (Symposium on Applied Computing)*, and some others on specific workshops associated to known conferences. We also published seven papers in the *Brazilian Symposium on Databases*, the main national database conference, which has an international program committee, a low acceptance rate (< 25%) and is indexed by DBLP. All the publications are listed hereafter in chronological order.

[Silva et al. 2008] Silva, J., Fidalgo, R. N., Salgado, A. C., Times, V. C. Modelling and Querying Geographical Data Warehouses, *Information Systems*, edited by Elsevier, 2008 (submitted to evaluation)

[Silva et al. 2007] Silva, J., Fidalgo, R. N., Oliveira, A., Salgado, A. C., Times, V. C. Querying Geographical Data Warehouses With GeoMDQL, In Proc of the *22nd Brazilian Symposium on Databases*, João Pessoa, Brazil, 2007. p.223-237

[Silva et al. 2006] Silva, J., Times, V. C., Salgado, A. C., Medeiros, V. N., Fidalgo, R. N.An Open Source and Web-based Framework for Geographic and Multidimensional Processing, In Proc of the *21st Annual ACM Symposium on Applied Computing*, Dijon, France, 2006, p.63-67

[Souza et al. 2006] Souza, D., Salgado, A. C., Tedesco, P. C. A. R. Towards a Context Ontology for Geospatial Data Integration, In Proc of the *International Workshop on Semantic-based Geographical Information Systems*, LNCS 4278, Montpellier, France, 2006, p.1576-1585

[Silva et al. 2004] Silva, J., Times, V. C., Salgado, A. C., Fidalgo, R. N. Propondo uma Linguagem de Consulta Geográfica Multidimensional, In Proc of the *6th Brazilian Symposium on Geoinformatics*, Campos do Jordão, Brazil, 2004. p.479-489

[Fidalgo e al. 2004] Fidalgo, R. N., Silva, J., Souza, F. F., Times, V. C., Salgado, A. C. Providing Multidimensional and Geographical Integration Based on a GDW and Metamodels, In Proc of the *19th Brazilian Symposium on Databases*, Brasília, Brazil, 2004. p.148-162

[Ciferri et al. 2003] Ciferri, R. R., Salgado, A. C., Times, V. C., Nascimento, M., Magalhaes, G. A Performance Comparison among Traditional R-Tree, Hilbert R-Tree and SR-Tree, In Proc of the *23rd International Conference of the Chilean Computer Science Society (SCCC2003)*, Chillán, Chile, 2003. p.3-12

[Soares & Salgado 2002] Soares, V. G., Salgado, A. C. Visual Query in Geographic Information Systems, In Proc of the *6th IFIP Working Conference On Visual Database Systems (VDB6)*, Brisbane, Australia, 2002, p.251-265

[Ciferri & Salgado 2001a] Ciferri, R. R., Salgado, A. C. Investigating the Performance Variation of Multidimensional Access Methods in Terms of Spatial Data Distribution. In *ACM SIGMOD Digital Symposium Collection*, Chicago, USA, 2001.

(http://www.acm.org/sigs/sigmod/disc/disc01/out/p_investigatingthrian.htm)

[Cifferi & Salgado 2001b] Ciferri, R. R., Salgado, A. C. Análise de Eficiência de Métodos de Acesso Espaciais em Termos da Distribuição Espacial dos Dados, In Proc of the *3rd Workshop Brasileiro de Geoinformática*, Rio de Janeiro, Brazil, 2001, p.79-86

[Beltrão et al. 2001] Beltrao, V. P. C. R., Times, V. C., Salgado, A. C. Guia Metropole: Um Sistema de Roteamento de Veículos Utilizando um SIGWEB, In Proc of the *Conferencia Latino Americana de Informática*, Mérida, Venezuela, 2001, p.30-38

[Perez & Salgado 2001a] Perez, C. R., Salgado, A. C. Sistema de Informações Geográficas Hipermídia Aberto, In Proc of the *9th Encuentro Chileno de Computación*, Punta Arenas, Chile, 2001.

[Perez & Salgado 2001b] Perez, C. R., Salgado, A. C. Uma Arquitetura para Integração e Interoperabilidade de Sistemas de Informações Geográficas, In Proc of the *7th Argentina Congress on Computer Science*, El Calafate, Argentina, 2001.

[Soares & Salgado 2000] Soares, V. G., Salgado, A. C. A Metadata-based Approach to Define a Standard to Visual Queries in GIS, In Proc of the *International Workshop on Interacting with Database (in conjunction with DEXA)*, Greenwich, England, 2000, p.693-697

[Ciferri & Salgado 2000a] Ciferri, R.R., Salgado, A.C. Performance Evaluation of Multidimensional Access Methods, In Proc of the *8th ACM Symposium on Advances in Geographic Information Systems*, Washington D.C., 2000, p.183-184

[Cifferi & Salgado 2000b] Ciferri, R. R., Salgado, A. C. Fatores Determinantes de Desempenho de Métodos de Acesso Multidimensionais, In Proc of the *2nd Workshop Brasileiro de GeoInformática (GeoInfo2000)*, Sao Paulo, Brazil, 2000, p.112-119

[Souza & Salgado 2000a] Souza, D., Salgado, A. C. Aplicacao de Fatores Multidisciplinares na Construcao de uma Interface de Consulta para Sistemas de Informações Geográficas, In Proc of the *3rd Workshop sobre Fatores Humanos em Sistemas Computacionais*, Gramado, Brazil, 2000

[Souza & Salgado 2000b] GeoVisual Interface - A Visual Query Interface for Geographic Information Systems, In Proc of the *15th Brazilian Symposium on Databases*, João Pessoa, Brazil, 2000, p.7-19

[Ciferri et al. 2000] Ciferri, R. R., Cortes, S.S., Salgado, A. C. Investigando a Variação de Desempenho de MAM em Função de Distribuição Espacial de Dados, In Proc of the *15th Brazilian Symposium on Databases*, Joao Pessoa, Brazil, 2000. p.115-129

[Soares & Salgado 1999] Soares, V. G., Salgado, A. C. Consultas Visuais em Sistemas de Informações Geográficas baseadas em Padrões de Metadados Espaciais, In Proc of the *1st Workshop Brasileiro de GeoInformática (GeoInfo1999)*, Campinas, Brazil, 1999, p.14-23

[Perez & Salgado 1999a] Perez, C. R., Salgado, A. C. SIGHA: Uma Arquitetura Aberta e Interoperável para Sistemas de Informações Geográficas, In Proc of the *26th SEMISH - Congresso da SBC*, Rio de Janeiro, Brazil, 1999. p.135-148

[Perez & Salgado 1999b] Perez, C. R., Salgado, A. C. Sistemas Hipermídia Abertos: Uma Solução ao Problema de Interoperabilidade e Integração de Sistemas, In Proc of the *25th Conferencia Latino-Americana de Informática*, Assunção, Paraguai, 1999, p.25-37

[Perez et al. 1997] Perez, C. R., Ferraz, C. A. G., Salgado, A. C. Processamento de Informações Geográficas Distribuídas: Arquiteturas para as Redes de Serviços Públicos, In Proc of the *23rd Conferência Latino-Americana de Informática*, Valparaiso, Chile, 1997.

[Nascimento & Salgado 1995] Nascimento, A. R., Salgado, A. C. Lingeo-Uma Linguagem de Consulta Geográfica, In Proc of the *10th Brazilian Symposium on Databases (SBBD)*, Recife, Brazil, 1995, p.147-161

[Cavalcanti & Salgado 1994] Cavalcanti, A. E. C., Salgado, A. C. Um Estudo para Tratar a Dimensão de Tempo em Sistemas de Banco de Dados, In Proc of the *9th Brazilian Symposium on Databases (SBBD)*, São Carlos, Brazil, 1994. p.357 - 381

[Times & Salgado 1994] Times, V., Salgado, A.C. Object-oriented Modeling for Geographical Applications, In Proc of the *9th Brazilian Symposium on Databases (SBBD)*, São Carlos, Brazil, 1994, p.293-309

2.6 References

[Beckmann et al. 1990] Beckmann, N., Kriegel, H.-P., Schneider, R., and Seeger, B. The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles, In Proc. of the *ACM SIGMOD International Conference on Management of Data*, Atlantic City, USA, 1990, p. 322-331

[Berman & Stonebraker 1977] Berman, R.R., Stonebraker, M. GEO-QUEL - A System for the Manipulation and Display of Geographic Data, In Proc. of the *4th Annual Conference on Computer Graphics and interactive Techniques*, San Jose, USA, 1977, p. 186-191

[Bowen 1992] Bowen, S., Z Bibliography, *Z Technical and Users Meeting*, England, 1992

[Burrough & McDonnell 1998] Burrough, P., McDonnell, R. Principles of Geographical Information Systems, *Oxford University Press*, 1998.

[Calcinelli & Mainguenaud 1994] Calcinelli, D., Mainguenaud, M. Cigales: A Visual Query Language for Geographical Information System: The User Interface. *Journal of Visual Languages and Computing*, Vol. 5, N.2, 1994, p.113-132

[Carneiro & Magalhães 1998] Carneiro, A.P., Magalhães, G.C. Usando Dados Reais Urbanos na Comparação de Desempenho de Métodos de Acesso Espaciais, In Proc of the *13th Brazilian Symposium on Databases*, Maringá, Brazil,1998, p.401-415

[Catarci et al. 1997] Catarci, T., et.al. Visual Query Systems for Databases: Analysis and Comparison. *Journal of Visual Languages and Computing*, Vol. 8, N. 2, 1997, p. 215-260.

[Chang & Fu 1980] Chang, N.S., Fu, K.S. Query-by-Pictorial-Example, *IEEE Transactions on Software Engineering*, Vol.6, N.6, 1980, p. 519-524

[Cox 1991] Cox Junior, F.S. Análise de Métodos de Acesso a Dados Espaciais Aplicados a Sistemas Gerenciadores de Banco de Dados. *Master's thesis*, DCC, Unicamp, Campinas, Brazil, 1991, 171 pp

[Danko 1997] Danko, D.M. Perspectives in the Development of ISO Metadata Standards, In Proc. of the *Earth Observation World Wide Web Workshop*, Washington, USA, 1997

[Dueker & Kjerne 1987] Dueker, K., Kjerne, D. Application of the Object-Oriented Paradigm to problems in Geographic Information Systems. In Proc. of the *International GIS Symposium: The Research Agenda*, 1987, p.79-88

[Egenhofer & Frank 1987] Egenhofer, M.J., Frank, A.U. Object-Oriented Databases: Database Requirements for GIS. In Proc. of the *International GIS Symposium: The Research Agenda*, 1987, p.189-211

[Egenhofer 1993] Egenhofer, M.J. What's Special about Spatial Database Requirements for Vehicle Navigation in Geographic Space, In Proc. of the *ACM SIGMOD International Conference on Management of Data*, Washington, D.C., USA, 1993, p.398-402

[Egenhofer 1994] Egenhofer, M.J. Spatial SQL: A Query and Presentation Language. *IEEE Transaction on Knowledge and Data Engineering*, Vol. 6, N.1, 1994, p.86-95

[FGDC 1999] FGDC Standards Reference Model, *Federal Geographic Data Committee*, 1999, http://www.fgdc.gov/

[Frank 1982] Frank, A. MAPQUERY: Database Query Language for Retrieval of Geometric Data and their Graphical Representation**,** *ACM SIGGRAPH Computer Graphics*, Vol.16, N.3, 1982, p. 199-207

[Gaede & Günther 1998] Gaede, V., Günther, O. Multidimensional Access Methods, *ACM Computing Surveys*, Vol.30, N.2, 1998, p. 170-231

[Gardarin 1993] Gardarin, G. Integrating classes and relations to model and query geographical databases, In Proc. of the *4th International Conference on Database and Expert Systems Applications (DEXA)*, Prague, Czechoslovakia, 1993, p. 365-372

[Golin 1991] Golin, E. J. A Method for the Specification and Parsing of Visual Languages. *PhD Dissertation*, Brown University, Technical Report No. CS-90-19, 1991.

[Golin & Reiss 1990] Golin, E.J., Reiss, S.P. The Specification of Visual Language Syntax. *Journal of Visual Languages and Computing*, Vol.1, N.2, p. 141-157.

[Güting 1994] Güting, R.H. An Introduction to Spatial Database Systems, *The VLDB Journal*, Vol.3, N.4, 1994, p. 357-399

[Guttman 1984] Guttman, A. R-Trees: A Dynamic Index Structure for Spatial Searching, In Proc. of the *ACM SIGMOD International Conference on Management of Data*, Boston, USA, 1984, p. 47-57

[Joseph & Cardenas 1988] Joseph, T., Cardenas, A.F. PICQUERY: a high level query language for pictorial database management, *IEEE Transactions on Software Engineering*, Vol.14, N.5, 1988, p. 630-638

[Kriegel et al. 1989] Kriegel, H.-P., Schiwietz, M., Schneider, R., and Seeger, B. Performance Comparison of Point and Spatial Access Methods, In Proc. of the *1st International Symposium on the Design and Implementation of Large Spatial Databases(SSD)*, LNCS 409, 1989, p. 89-114

[Lapalme et al. 1992] Lapalme, G., Rousseau, J-M, Chapleau, S., Cormier, M., Cossette, P., Roy, S. A Geographic Information Systems for Transportations Applications. *Communications of the ACM*, Vol.35, N.1, 1992, p.80-88

[Meira & Cavalcanti 1992] Meira, S.R.L., Cavalcanti, A.L.C. The MooZ Specification Language, *Techical Report ES/1.92*, UFPE, 1992

[Meyer 1992] Meyer, B. Beyond Icons – Towards New Metaphors for Visual Query Languages for Spatial Information Systems. *Interfaces to Database Systems*, R. Copper (Ed.), Springer, 1992

[OGC 1999] Open GIS Consortium, Simple Features Specification for SQL, Revision 1.1., 1999, http://www.opengis.org/

[Oliveira & Medeiros 1995] Oliveira, J. L., Medeiros, C. A Direct Manipulation User Interface for Querying Geographic Databases. In Proc. of the 2nd *International Conference Applications on Databases*, San Jose, USA, 1995, p. 249-258

[Ooi et al. 1989] Ooi, B.C., Sacks-Davis R., McDonell, K.J. Extending a DBMS for Geographic Applications, In Proc. of the *5th International Conference on Data Engineering (ICDE)*, Los Angeles, USA, 1989, p.590-597.

[Oosterom & Bos 1989] Oosterom, P.V., Bos, J.V.D. An Object-Oriented Approach to the Design of Geographic Information Systems, *Computers & Graphics*, Vol.13, N.4, 1989, p.409-418

[Orenstein & Manola 1988] Orenstein, J.A., Manola, F.A. PROBE Spatial Data Modeling and Query Processing in an Image Database Application, *IEEE Transactions on Software Engineering*, Vol.14, N.5, 1998, p.611-629

[Peuquet 1984] Peuquet, D.J. A Conceptual Framework and Comparison of Spatial Data Models, *Cartographica*, Vol.21, N.4, 1984, p.66-113

[Rumbaugh et al. 1991] Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., Lorensen, W. *Object-Oriented Modeling and Design*, Prentice-Hall International Editions, 1991.

[Rumbaugh et al. 1999] Rumbaugh, J., Booch, G., Jacobson, I. The Unified Modeling Language – User Guide, Addison-Wesley, 1999

[Roberts et al. 1991] Roberts, S.A., Gahegan, M.N., Hogg, J., Hoyle, B. Application of Object-Oriented Databases to Geographic Information Systems, *Information and Software Technology*, Vol.33, N.1, 1991, p.38-46

[Roussopoulos & Leifker 1985] Roussopoulos, N., Leifker, D: Direct spatial search on pictorial databases using packed R-trees, In Proc. of the *ACM SIGMOD International Conference on Management of Data*, Austin, USA, 1985, p.17-31.

[Sacks-Davis et al. 1987] Sacks-Davis, R., McDonell, K.J., Ooi, B.C. GEOQL – A Query Language for Geographic Information Systems, *Technical Report 87/2*, Monash University, Australia, 1987

[Sellis et al. 1987] Sellis, T., Roussopoulos, N., and Faloutsos, C. The R^+-Tree: A Dynamic Index for Multi-Dimensional Objects, In Proc. of the *13th International Conference on Very Large Data Bases (VLDB)*, Brighton, England, 1987, p. 507-518

[Shneiderman 1998] Shneiderman, B. Designing the User Interface: Strategies for Effective Human-Computer Interation, Addison Wesley, 1998.

[Stonebraker & Rowe 1986] Stonebraker, M., Rowe, L.A. The Design of POSTGRES, In Proc. of the *ACM SIGMOD International Conference on Management of Data*, Washington, USA, 1986, p. 340-355

[Williamson & Stucky 1992] Williamson, R., Stucky, J. An object-oriented geographical information system, *Object-Oriented Databases with Applications to Case, Networks and VLSI CAD*, Prentice-Hall Series in Data and Knowledge Base Systems, editors R. Gupta and E. Horowitz, 1991, p.296-312.

CHAPTER 3

Data Integration Systems

3.1 Introduction

The data integration systems are tools that offer a uniform access to distributed and heterogeneous Web data sources. This is done by resolving the heterogeneities and giving to the disparate sources an integrated view, i.e., a collection of views over the data sources reflecting the users' requirements. Users submit queries over the integrated view without having to spend a lot of time in searching and browsing the Web.

Classical systems are based on two approaches to data integration, each one with a specific implementation architecture [Abiteboul et al. 1999]:

Virtual approach — in this approach, the data remains in the sources and queries submitted to the data integration system are decomposed into queries addressed directly to the sources. In virtual data integration systems [Chawathe et al. 1994], a software module, called mediator, receives a query, decomposes it into sub-queries over the data sources and integrates its results.

Materialized approach — in this approach, data are previously accessed, cleaned, integrated and stored in a data warehouse [Widom, 1995] and the queries submitted to the integration system are evaluated in this repository without direct access to the data sources.

Many data integration systems use the mediation architecture [Wiederhold 1992] to provide integrated access to multiple data sources. Such mediation systems can be classified according to the approach used to define the mediation mappings between the data sources and the global schema [Halevy 2000, Ullman 1997]. The first approach is called Global-As-View (GAV) and requires that each object of the global schema be expressed as a view (i.e., a query) on the data sources. In the other approach, called Local-As-View (LAV), mediation mappings are defined in an opposite way; each object in a given source is defined as a view on the global schema. The Global-Local-as-View (GLAV) [Friedman et al. 1999, Lenzerini 2002] and Both-as-View (BAV) [McBrien & Poulovassilis 2003] approaches combine features of both GAV and LAV approaches.

Several data integration systems have been proposed in the literature with different approaches. Some of them are TSIMMIS [Chawathe et al. 1994], Nimble [Draper et al. 2001], WHIPS [Labio et al. 1997], MOMIS [Bergamaschi et al. 1998] and e-XML [Gardarin et al. 2002]. In what follows we present *Integra*, our proposal to a data integration system, along with its architecture, the X-Entity model, the definition and evolution of mediation queries, the quality evaluation of the generated mediation schema and the results obtained.

In this research area, two projects had financial support from Brazilian institutions (CNPq and FINEP) from 2003 to 2006. Currently, an international project is running with France and Uruguay (AMSUD-STIC).

3.2 The Integra System

Integra is an information integration system which adopts the GAV approach and uses XML [Bray et al. 2006] as a common model for data exchange and integration [Loscio 2003]. It combines features of both data integration approaches supporting the execution of virtual and materialized queries. Some portions of data more intensively unavailable and static may be materialized in a data warehouse [Amaral 2007] and the more dynamic data are accessed by virtual queries [Batista et al. 2003, Batista 2003].

Another distinguishing feature of our approach is the use of a local cache, i.e., a repository to store prepared answers for the most frequently queries submitted to the integration system [Galvao 2007]. The cache idea is to immediately return results for some user queries with minimal processing time. The key issue of our work consists in the creation of a data integration environment that supports three kinds of queries and returns the respective data to answer them: (i) Virtual data which are obtained on demand and accessed directly from the data sources; (ii) Materialized data which are obtained from the data warehouse over selectively materialized data; (iii) Cached queries which are answered by the retrieval of ready results previous stored in a local cache.

As shown in Figure 3.1, the system architecture can be divided into four spaces:

Common core: this space feeds the mediator generation and maintenance space with information about local data source schemas while receiving local data source queries from the data integration space and answering them.

Data integration space: the main component of this space is the mediator which is responsible for restructuring and merging data from autonomous data sources and for providing an XML integrated view of the data. Other components of this space are used to optmize the overall query response time of user queries.

Mediator generation and maintenance space: through semi-automated processes this space executes the mediation queries generation and maintenance. The process of mediation queries generation is based on the approach for discovering relational view expressions proposed by Kedad [Kedad & Bouzeghoub 1999]. Since we adopt XML as the common data model, we had to adapt this approach in order to generate XML-based mediation queries. The mediation queries maintenance is executed by a global evolution process, which receives events about data source schemas changes and users' requirements changes, and propagates them into the mediation level [Loscio & Salgado 2004].

User space: the components of this space are used to specify the user requirements and to manage their evolution. The user requirements are used as input to the process of mediation schema definition. The evolution of user requirements are propagated to the mediation schema and the mediation queries through the components responsible for the mediation queries evolution.

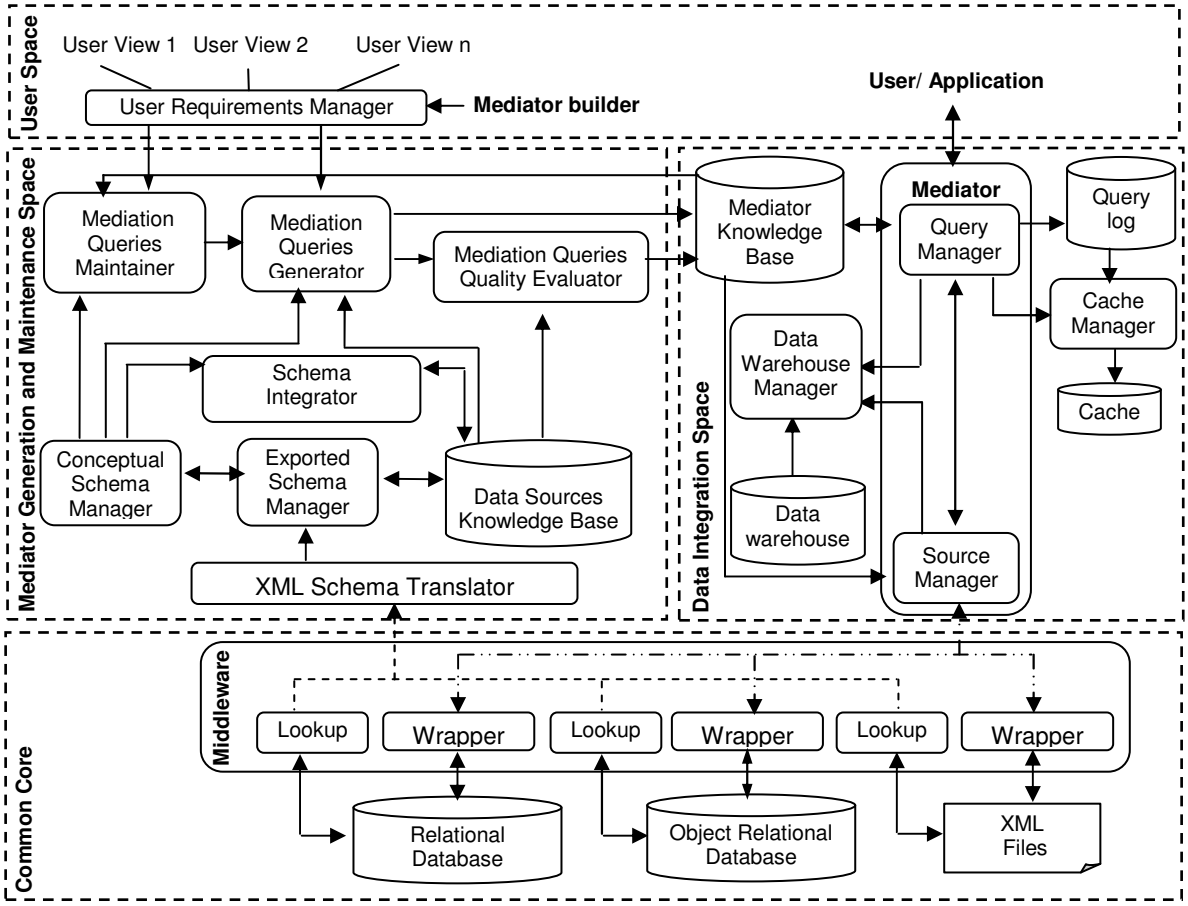


Figure 3.1 – *Integra* Architecture overview

In this proposed environment the main issues addressed and presented hereafter were:

i. The proposal of X-Entity: a conceptual model to provide a high-level abstraction for information described in an XML Schema [Fallside & Walmsley 2004]. Although being very useful for validating XML documents, an XML Schema is not suitable for tasks requiring knowledge about the semantics of the represented data.

ii. The definition and maintenance of mediation queries: since the GAV approach is adopted, the system must provide a uniform view of the underlying data sources, called mediation schema, and must define a set of mediation queries which compute each object in the mediation schema. One of the main problems in this context is the maintenance of the mappings between the mediation and the source schemas. Each change at the source schema level may lead to the reconsideration and possibly the change of all mediation queries.

iii. The reformulation of user queries: one of the challenges facing a data integration system consists in answering a user query submitted in terms of the mediation schema given that the data is at the sources. This problem consists mainly in reformulating the query in terms of the source schemas and in integrating the corresponding answers.

iv. The quality of the integrated schema: the main issue is to incorporate Information Quality analysis into data integration systems, particularly in the integrated schema. The adopted criteria were schema minimality, type consistency and completeness.

3.3 The X-Entity Model

X-Entity model [Loscio et al. 2003] was proposed to provide a high-level abstraction for information described in an XML Schema [Fallside & Walmsley 2004]. The X-Entity model is not a new formalism for conceptual modeling, rather it is an extension of the Entity Relationship model [Chen 1976], i.e., it uses some basic features of the ER model and extends it with some additional ones to better represent XML Schemas. Each element is seen and manipulated as an individual concept even if it belongs to a nested structure. Instead of having nested elements, each element has a set of relationships that represent its association with other elements. In this sense,, the X-Entity model provides support for the data integration system offering a concise and semantic description for XML Schemas.

3.3.1 X-Entity concepts

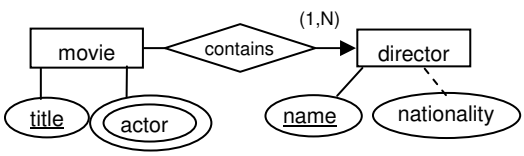
An X-Entity schema S is denoted by $S = (E, R)$, where E is a set of entity types and R is a set of relationship types.

Entity type: an entity type E, denoted by $E(\{A_1,\ldots,A_n\},\{R_1,\ldots,R_m\})$, is made up of an entity type name E, a set of attributes $A_1,\ldots,A_n$ and a set of relationship types $R_1,\ldots,R_m$. An entity type represents a set of elements with a complex structure, composed by attributes and other elements (called subelements). An instance of an entity type represents a particular element in the source XML document. Each entity type has attributes $\{A_1,\ldots,A_n\}$ that describe it. An attribute A_i represents either an attribute or a subelement, which is not composed by other elements or attributes. Each attribute A_i is associated with a domain, denoted $Dom(A_i)$, which specifies its value set. A_i is also associated with a cardinality, denoted $Card(A_i)=(min, max)$, which specifies the minimum and the maximum number of instances of A_i that can be related to an instance of E. In X-Entity diagrams, entity types are displayed as rectangles.

Containment relationship type: a containment relationship type between two entity types E_1 and E_2, specifies that each instance of E_1 contains instances of E_2. It is denoted by $R(E_1, E_2,(min,max))$, where R is the relationship name and (min, max) defines the minimum and the maximum number of instances of E_2 that can be associated with an instance of E_1. Each entity type E may be associated with one or more containment relationships, which describes the element-subelement relationship between E and other entity types. In X-Entity diagrams, containment relationships are displayed as diamond-shaped boxes labeled with contains. The straight lines connecting the relationship with the participating entities are directed from the entity E to the entity E1.

Reference relationship type: a reference relationship between two entity types E_1 and E_2, denoted by $R(E_1, E_2,\{A_{11},\ldots,A_{1n}\}, \{A_{21},\ldots,A_{2n}\})$, specifies that the entity type E_1 references the entity type E_2. $\{A_{11},\ldots,A_{1n}\}$ and $\{A_{21},\ldots,A_{2n}\}$ represent the referencing attributes between entities of E_1 and E_2 such that the value of A_{1i}, $1\leq i \leq n$, in any entity of E_1 must match a value of A_{2i}, $1\leq i \leq n$, in some entity of E_2. In X-Entity diagrams reference relationships are represented as a diamond-shaped box labeled with refers. The straight lines connecting the relationship with the participating entities are directed to the referenced entity.

Figure 3.2 presents an example of an X-Entity schema. The entity type *movie* has two attributes: *title* and *actor*. The attribute *title* is a key attribute of the entity type *movie*. The attribute *actor* is multivalued, which means that an element *movie* may have multiple occurrences of the subelement *actor*. In this example, there is one containment relationship, which specifies that an instance of *movie* has at least one subelement *director* and may have unlimited occurrences of the subelement *director*.



Schema of data source *S*:
movie({title,,actor},{movie_director})
director({name, nationality}, {})
movie_director(movie, director,(1,N))

Figure 3.2 - Example of an X-Entity schema

Since XML may be used to represent both structured and semi-structured data, X-Entity schemas are flexible, i.e., instances of the same entity type may have different structures. To represent this, both attributes and containment relationships are associated with participation constraints, which represent the occurrence constraints of elements and attributes in an XML document.

3.3.2 X-Entity Query Language (XEQ)

Queries submitted to a data integration system are, in most cases, defined using declarative query languages or templates. The great majority of such languages are intended to human comprehension, which leads to difficulties in the query decomposition [Deutsch & Tannen 2003] and translation [Hammer et al. 1997]. In order to minimize the complexity of such problems we proposed XEQ (X-Entity Query language), an XML-based language, used to provide an internal query representation for the *Integra* system [Costa 2005].

Considering that the queries submitted to the Integra system are specified according to a mediation schema defined in the X-Entity model, a XEQ expression uses X-Entity concepts to represent the queried data. The simplest conceptual representation of a XEQ expression involves naming one entity (i.e. the parent entity), providing a list of attributes to be returned, as well as using an expression to constrain this entity. We may also define expressions that return values for subentities of the parent entity. A subentity of an entity type *E* is an entity type *E'* which is associated with *E* through a containment relationship. Queries may associate parent entities and their referenced entities. A referenced entity type of an entity type *E* is an entity type *E'* which is associated with *E* through a reference relationship.

The basic element of a XEQ expression is the *SELECT* element. The *SELECT* element has one attribute, called *ENTITY*, whose value is the name of the entity type *E* being queried. The *SELECT* element may be composed by other elements, as: i) a *REFERENCE* element that defines the join condition used to combine related elements from the entity type *E* and a referenced entity type *E'*, ii) *ATTRIBUTE* elements that have in their content the name of an attribute of the entity type *E* whose value is to be retrieved by the query, iii) a *WHERE* element that filters the data to be retrieved by the query. To illustrate the use of the XEQ internal language we present, in Figure 3.3, an XQuery query and its corresponding XEQ expression.

During the translation to XEQ the query is parsed and each one of the entities being queried is identified and represented using a *SELECT* element. In this example, $book_2$ is the entity whose values should be retrieved. Since $chapter_2$ is a subentity of $book_2$, to retrieve such information it should be used a nested *SELECT* element. It is important to observe that in the *SELECT* element associated with the referenced entity $publisher_2$, there is a *MAP* element. This element indicates how the join operation between the entity $book_2$ and the entity $publisher_2$ should be done in order to retrieve the value of the $publisher\_name_2$ attribute.

User query: Retrieve the title, year, the title of the chapters and the publisher name of all books.

XQuery

```
<books> {
 for $b in //book2
 return
  <book2>
   <title2>$b/title2 </title2>
   <year2> $b/year2 </year2>
   <chapter2> $b/chapter2/chapter_title2 </chapter2>
    for $p in //publisher2
    where $b/ref_publisher2 =  $p/id_publisher2
    return
      <publisher2> $p/publisher_name2 </publisher2>
  </book2>
}</books2>
```

XEQ

```
<EXP>
 <SELECT ENTITY="book2">
  <ATTRIBUTE>title2</ATTRIBUTE>
  <ATTRIBUTE>year2</ATTRIBUTE>
   <SELECT ENTITY="chapter2">
     <ATTRIBUTE> chapter_title2 </ATTRIBUTE>
   </SELECT>
   <SELECT ENTITY="publisher2">
     <ATTRIBUTE> publisher_name2 </ATTRIBUTE>
     <REFERENCE NAME="ref_book_publisher">
       <MAP ATTRIBUTE="ref_publisher2"> id_publisher2 </MAP>
     </REFERENCE>
   </SELECT>
 </SELECT>
</EXP>
```

Figure 3.3 - User query example

3.3.2 Summary

X-Entity is a conceptual data model for XML schemas. Such representation provides a cleaner description for XML schemas hiding implementation details and focusing on semantically relevant concepts. With this representation one can explicitly represent important features of XML schemas, including: element and sub-element relationships, occurrence constraints of elements and attributes and choice groups. It is the basis for the formalization of mediation queries definition and evolution [Loscio 2003].

XEQ, an XML-based language, was defined to be used as an internal specification to facilitate both query decomposition and translation processes [Costa 2005].

3.4 Mediation Queries

A mediation query describes how to compute an element of the mediation schema over the data sources. Therefore, the process of mediation queries generation consists in discovering a computing expression for each entity in the mediation schema. In what follows we describe the mediation queries definition and the evolution of these mediation queries when data source schemas change.

3.4.1 Mediation Queries Definition

As the mediation schema is represented by an X-Entity schema, the process of mediation queries generation consists in discovering a computing expression for each entity in the mediation schema. More formally, we can say that defining a mediation query for a mediation entity *Em* consists in decomposing *Em* into n entity types *Ep1,...,Epn* such that *Em* = *Ep1*θ1*Ep2*θ2...θn*Epn*,

where θi is a binary operator and each entity type *Epi* is derived using an expression Exp(*E*i) over a single source entity type *Ei*. An entity type *Ei* is called a relevant source entity to compute *Em* and an entity type *Epi* is called a mapping view (m-view).

At the end of the mediation queries generation process, each mediation entity will be associated with a mediation query, which is represented by an operation graph. The operation graph describes all information that is relevant to compute a given integrated view. Another important issue to be considered is that an operation graph can be incrementally created and it can be easily modified.

The use of operation graphs to represent mediation queries facilitates the identification of the mediation entities that are affected by a data source schema change and that, consequently, should be rewritten. A mediation entity will be affected by an entity source change if one of its mapping views will be affected by the data source change. Consequently, the problem of propagating the source changes or users' requirements changes into the mediation queries consists, first in propagating these changes into the mapping views, and second in modifying the mediation queries in order to take into account the modifications in the set of mapping views.

To illustrate the mediation queries definition consider the mediation schema and the source schemas presented as follows.

Mediation Schema S_{med} =
$(\{movie_m(\{title_m, director_m\}, \{movie_m\_actor_m\}),$
$actor_m(\{name_m, nationality_m\},\{\})\},$
$\{movie_m\_actor_m(movie_m, actor_m(1,N))\})$

Schema of data source S_1 =
$(\{movie_1(\{title_1, duration_1\},$
$\{movie_1\_actor_1, movie_1\_director_1\}),$
$actor_1(\{name_1, nationality_1\},\{\})$
$director_1(\{name_1, nationality_1\},\{\})\},$
$\{movie_1\_actor_1(movie_1, actor_1,(1,N)),$
$movie_1\_director_1(movie_1, director_1, (1,N))\})$

Schema of data source S_2 =
$(\{movie_2(\{title_2, actor_2\},\{movie_2\_director_2\}),$
$director_2(\{name_2, nationality_2\},\{\})\},$
$\{movie_2\_director_2(movie_2, director_2,(1,N))\})$

Schema of data source S_3 =
$(\{director_3(\{name_3, nationality_3\},$
$\{director_3\_movie_3\}),$
$movie_3(\{title_3, year_3\},\{\}),$
$\{director_3\_movie_3(director_3, movie_3,(1,N))\})$

Figure 3.4 presents an example of an operation graph that describes the possible operators to combine the mapping views associated with the mediation entity $movie_m$. The nodes of the operation graph represent the mapping views (V_{Movie1}, V_{Movie2} and V_{Movie3}) and the edges between these mapping views represent the mapping operators. At the end of the mediation queries generation process, the mediation query $Q(movie_m) = (E_{pMovie1} \cup E_{pMovie3}) \cup E_{pMovie2}$ is obatined from the operation graph G_{moviem}. More details about the mediation queries generation process can be found in [Loscio 2003].

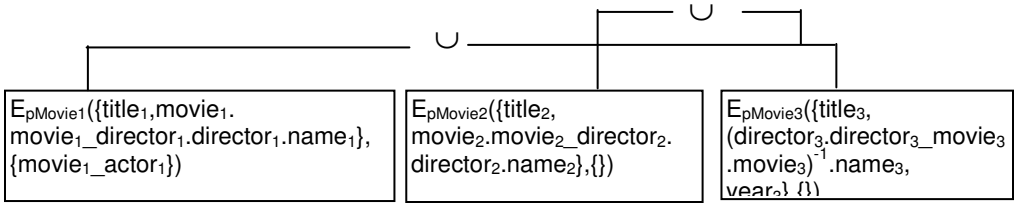


Figure 3.4 - Operation graph G_{moviem}

3.4.2 Mediation Queries Evolution

In a dynamic environment, the mappings between the mediation schema and the source schemas must be flexible enough in order to accommodate new data sources and new users' requirements. Each change at the source schema level may lead to the reconsideration and possibly the change of all mediation queries. The problem of mediation queries evolution is discussed in some other works. The work presented by Ambite et al. [Ambite et al. 2001] adopts an approach similar to ours for defining mediation queries. The algorithm to discover integration axioms is incremental, which means that when new sources are added, the system can efficiently update the axioms, but no details on how this could be achieved nor examples are given. In the case of deleting a source the algorithm must start from scratch. They use the LAV approach to define the mappings between the global model and the local sources. In [Mcbrien & Poulovassilis 2002] an approach is presented to handle both schema integration and schema evolution in heterogeneous database architectures. They use primitive transformations to automatically translate queries posed to the global schema to queries over the local schemas. In [Nica & Rundensteiner 1999] modifications are directly executed in the mediation query definition rather than in the metadata that describes the mediation query. Moreover, a view must evolve just when a source schema change makes the view definition obsolete, i.e., just the cases of removal of relations or attributes are dealt with.

In this context, we addressed a novel and complex problem that consists in propagating a change event occurring at the source level or at the user level into the mediation level. We propose an incremental approach to develop the mediation schema and the mediation queries based on the evolution of the data source schemas and the evolution of the users' requirements. The proposed approach allows the mediation level to evolve incrementally and modifications can be handled easier increasing the system flexibility and scalability. We deal with this problem by considering a specific context of data integration, where a mediation schema represents the reconciliation between users' requirements and the data sources' capabilities. Thus, users pose queries in terms of the mediation schema, rather than directly in terms of the source schemas.

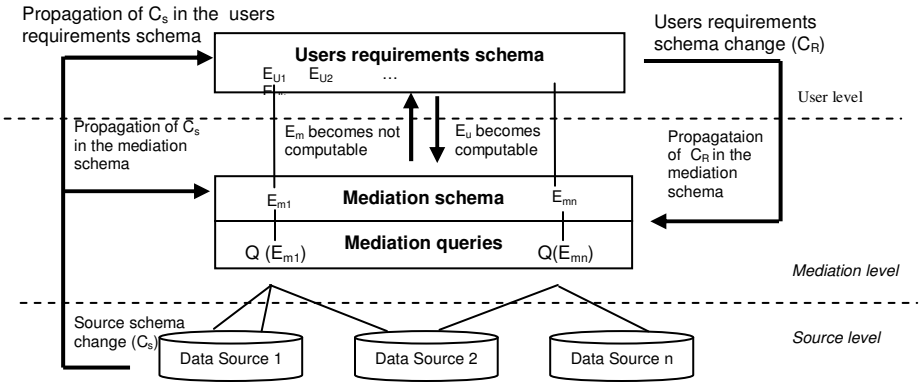


Figure 3.5- Propagation of data source schema changes and users' requirement changes

Figure 3.5 describes the impact of schema changes in the mediation level. As we may observe, the addition of new data sources and the modifications in the data source schemas are changes that must be propagated both to the user level and the mediation level. In the first case, the propagation involves an analysis of existing users' requirements to identify relevant source entities to compute user entities (E_u), i.e., entities participating in the users' requirements schema. The propagation of data source schemas changes to the mediation level consists mainly in changing the mediation queries when the changes raised at the source level still allow the computation of the mediation entities (E_m). However, some mediation entities may become no

longer computable concerning to the changes raised at the source level. Therefore, in these cases, the mediation entities must be removed from the mediation schema.

Concurrently with the evolution of the data source schemas, the users' requirements may continue to change. The evolution of the users' requirements schema originates directly change operations in the mediation schema. If these changes can be reflected in the mediation queries, the modifications on the mediation schema are committed; otherwise, the user is informed that his or her new requirements cannot be satisfied. This occurs when the requirement is not computable from the data available in the data sources. Each change raised in the mediation schema may lead to the redefinition of some mediation queries or the generation of new ones.

Data source schemas or users' requirements changes are propagated to the mediation queries through a set of event-condition-action (ECA) rules, which are triggered according to the different schema changes. The propagation process consists of two main tasks: first, the triggering, evaluation and execution of the rules in order to update the mapping views and the operations among them, and secondly, new mediation queries are generated using the modified operation graphs.

Propagation primitives

We consider that each mediation entity E_m is associated with an operation graph $G_{Em}(M_{Em},O_{Em})$ corresponding to the mediation query defining E_m, where M_{Em} is the set of nodes of G_{Em}, representing the set of m-views associated with E_m, and O_{Em} is the set of edges of G_{Em}, labeled with one of the mapping operators. If a change occurs in the data source schemas or in the mediation schema, some checking operations have to be performed on this graph to test if the mediation query associated with E_m are still valid. If not, a new mediation query has to be defined. Each entity E_m in the mediation schema is associated with two attributes, *MAPSET_STATUS* and *OPSET_STATUS*. These attributes have boolean values and represent the status of the set of mapping views associated with $E_m(M_{Em})$ and the set of candidate operations to combine these entities (O_{Em}). They determine if the set of mapping views associated with E_m (M_{Em}) and the set of candidates operations to combine these views (O_{Em}) were modified during the propagation of the schema changes. These two attributes are set to *False* at the beginning of the propagation process, and they will be set to *True* if a change occurs in the set of mapping views or the set of operations respectively. The set of propagation primitives is presented in Table 3.1. These primitives are used in the mapping view evolution rules.

Table 3.1 - Mediation queries propagation primitives

Mediation Level Propagation Primitive	Definition
search_operation(G_{Em})	Searches new operations for combining pairs of mapping views in the operation graph G_{Em}. If new operations are generated, then the attribute *OPSET_STATUS* is set to *TRUE*.
remove_operations(G_{Em},V,A)	Removes all edges in the operation graph G_{Em} that become invalid because of the removal of the attribute A from the mapping view V. If at least one operation is removed, then the attribute *OPSET_STATUS* is set to *TRUE*.
add_mapping(V, G_{Em})	Adds a mapping view V into the operation graph G_{Em} and assigns the *TRUE* value to the attribute *MAPSET_STATUS*. If G_{Em} does not exist then this primitive creates G_{Em} from V.
remove_mapping(V, G_{Em})	Removes the mapping view V from the operation graph G_{Em} and assigns the *TRUE* value to the attribute *MAPSET_STATUS*.

Mapping views evolution rules

Given a change represented by one of the schema change operations described in section 3.2, we will first propagate these changes in the set of mapping views associated with each entity of the mediation schema. To specify this propagation, we use event-condition-action (ECA) rules. Due to space limitations, in this section, we present just some rules to illustrate the propagation of schema changes into mapping views. In the following, consider:

- *E_m: mediation entity*
- *V: mapping view*
- *E_i: source entity*
- *ADP: attribute derivation path*
- *EDP: entity derivation path*
- *G_{Em}: operation graph of the mediation entity E_m*
- *M_{Em}: is the set of mapping views corresponding to the mediation entity E_m*
- *$X(E_i)$: is the set of mapping attributes between the source entity E_i and the other source entities which originated mapping views belonging to M_{Em}*
- *$V = Exp(E_i)$: specifies that the mapping view V is derived from the source entity E_i*

Propagation of data source schemas changes

We propose a set of rules to propagate data source schemas changes into mapping views. As described before, a mapping view specifies how to compute attributes and subentities of a mediation entity E_m from a source entity E_i. A mapping view $V(\{X_1,...,X_n\},\{Y_1,...,Y_m\})$ is a special entity type where X_i is an attribute or an attribute derivation path and Y_i is a relationship or an entity derivation path. So, the propagation of a data source schema change into a mapping view V may result in the addition or removal of an attribute, relationship or derivation path from V. Each rule has a name and a parameter denoted E_m, which represents a mediation entity. To illustrate the propagation of a data source schema change consider the following rules (Rule 1 and Rule 2), which update the set of mapping views associated with the entity E_m in the mediation schema after the deletion of a source entity E_i. It is important to observe that it may exist more than one rule for each one of the source schema change operations.

Rule 1(E_m)

When remove_entity(E_i, S)

If $\exists V \in M_{Em} \mid V = Exp(E_i)$

Then remove_mapping(V, G_{Em})

Rule 1: the condition part of this rule checks if there is a mapping view V associated with E_m over E_i. To reflect the deletion of the local entity E_i, the corresponding mapping view V must be removed from the operation graph G_{Em}, along with all the operations involving the mapping view V. It is important to observe that just the mapping view corresponding to the source entity E_i is removed from the operation graph G_{Em}.

Rule 2(E_m)

When remove_entity(E_i, S)

If $\exists E_j \cong E_m \mid \exists V \in M_{Em} \wedge V = Exp(E_j)$

Then

If $\exists \{ADP_1, ...,ADP_n\}$, where $ADP_t = (E'.....E_i.R_k. ... E_j)^{-1}.A_k \mid$

$\forall t = 1,...,n, \exists E_m.A_m \cong ADP_t$ **or** $\exists \{EDP_1, ...,EDP_p\}$, where $EDP_t = (E'.....E_i.R_k. ... .E_j)^{-1} \mid$

$\forall t = 1,...,p, \exists E_m.R_m.E_m' \cong EDP_t$

Then $\underline{V.A} := \underline{V.A} - \{ADP_1,...,ADP_n\}$, $\underline{V.R} := \underline{V.R} - \{EDP_1,...,EDP_p\}$

remove_operations(G_{Em}, V, $\{ADP_1,...,ADP_n\}$)

Rule 2: the condition part of this rule checks if there is a source entity E_j that is semantically equivalent to E_m and if there is a mapping view V associated with E_m over E_j. In this case, the condition part of the rule also verifies if there are derivation paths from the entity type E_j that cannot be computed after the deletion of E_i. In this case, these derivation paths (attribute derivation path or entity derivation path) must be removed from the mapping view V. Then, it is necessary to verify if there are some mapping operators between the mapping view V and other mapping views V' which depend on one of the removed attribute derivation paths. In this case, the mapping operator must be removed.

Propagation of users' requirements changes

The propagation of users' requirements changes into mapping views is done through a set of mapping views evolution rules, which identifies information from the source entities relevant to the computation of the new requirements and performs the necessary modifications into the set of mapping views. Depending on the operation, the rules are evaluated over the whole set of source entities or over the mapping views associated with the mediation entity that is being modified. Each rule has a name and a parameter denoted V, which represents a mapping view or a parameter E_i which represents a source entity. To illustrate the propagation of a user requirement change consider the following rule (Rule 3), which identifies the mapping views relevant to compute a new mediation entity E_m. Rule 3 is evaluated for each source entity E_i. If E_i is semantically equivalent to E_m then the attributes, containment relationships and derivation paths of E_i, which are considered relevant to compute E_m, are identified. Such elements will compose the content of the new mapping view V, which will be inserted into the operation graph G_{Em}.

Rule 3(E_i)

When add_entity($\underline{E_m}$, S_m)

If $E_i \in \underline{S.E} \cong E_m$

Then if $\exists\{A_1,...,A_n\} \in \underline{E_i.A} \mid \forall t = 1,...,n, \exists E_m.A_m \cong E_i.A_t$ **or**

$\exists\{ADP_1, ...,ADP_n\}$,where $ADP_t = E_i. ... .A_k \vee ADP_t = (E'. ... .E_i)^{-1}.A_k \mid$

$\forall t = 1,...,n, \exists E_m.A_m \cong ADP_t$ **or**

$\exists\{R_1,...,R_k\} \in \underline{E_i.R} \mid \forall t = 1,...,n, \exists E_m.R_m.E_m' \cong E_i.R_t.E'$ **or**

$\exists\{EDP_1, ...,EDP_p\}$, where $EDP_t = E_i. ... .E' \vee EDP_t = (E'. ... .E_i)^{-1} \mid$

$\forall t = 1,...,n, \exists E_m.R_m.E_m' \cong EDP_t$

Then $\underline{V_{Ei}.A}:=\{A_1,...,A_n\}\cup\{ADP_1,..., ADP_n\}\cup X(E_i)$, $\underline{V_{Ei}.R} :=\{R_1,...,R_p\}\cup\{EDP_1, ...,EDP_p\}$

add_mapping($V_{\underline{Ei}}$, G_{Em}), search_operation(G_{Em})

As the structure of XML data is more flexible than the structure of conventional data, the process of propagating a schema change into a set of XML-based mediation queries is more complex. Such propagation process consists mainly in updating the attributes, containment relationships and reference relationships of relevant mapping views rather than just updating attributes. One advantage of our approach is that we use path correspondence assertions to capture the correspondences between elements with different structures.

3.4.2 Summary

We have presented the process of managing the evolution of XML-based mediation queries. Changes to mediation queries may be due to changes in the users' requirements or in data source schemas. The proposed solution was developed as part of the *Integra* data integration system which adopts the GAV approach. This work was part of Bernadette Loscio's PhD thesis [Loscio

2003, Loscio & Salgado 2004, Loscio et al. 2003, Bouzeghoub et al. 2003, Loscio & Salgado 2003, Loscio et al. 2002, Loscio et al. 2001].

3.5 Query Reformulation

Queries submitted to a data integration system are, in most cases, defined using declarative query languages or templates. The great majority of such languages are intended to human comprehension, which leads to difficulties in the query decomposition and translation. Another problem faced during query processing consists in translating data source queries to the native query language. To improve the query translation process, data integration systems use languages based on scripts or templates as the input format for wrappers [Hammer et al. 1997]. There are also systems that use XML query languages as input format for wrappers. The e-XML [Gardarin et al. 2002], MARS [Deutsch & Tannen 2003] and XPERANTO [Shanmugasundaram et al. 2001] systems provide integration of multiple data sources based on XML schemas. AutoMed [Boyd et al. 2004] adopts a functional language, called IQL, to provide a common query language where queries written in high level query languages can be translated into and out of. In order to minimize the complexity of such problems we proposed XEQ (X-Entity Query language), an XML-based language, used to provide an internal query representation for the *Integra* system [Costa 2005].

3.5.1 The Query Reformulation Process

Query execution consists in receiving a user query, expressed in a declarative query language, and in returning a query answer expressed in XML as output. The query execution process may be summarized as follows:

i) The mediator receives a user query Q and performs the necessary translation and reformulation to produce the set of local XEQ expressions (q_1, q_2, …, q_n),

ii) Next, the mediator sends these expressions to the corresponding wrappers which translate them to the source query language producing the local queries ($q_{1'}$, $q_{2'}$, …, $q_{n'}$),

iii) The wrappers receive answers for such queries ($r_{1'}$, $r_{2'}$, …, $r_{n'}$) from the data sources, these answers are translated to XML and sent to the mediator and

iv) Finally, the mediator integrates the XML answers (r_1, r_2, …, r_n) and returns the integrated result R as the answer to the original user query Q.

To describe the basic algorithm that reformulates a XEQ expression over the mediation schema into a set of XEQ expressions over the source schemas lets consider Q the original user query and *Xo* the expression obtained after the translation of Q to the XEQ representation. There are three phases in the algorithm:

1. Identification of the mediation entity (*Em*) being queried: this entity is specified in the most external select element of the *Xo* expression,
2. Identification of data sources relevant to answer the user query: a data source DSi is relevant to answer the user query Q if it contains one or more source entities necessary to compute the mediation entity *Em*, i.e., source entities semantically equivalent to *Em*,
3. Reformulation phase: this phase consists in generating a XEQ expression for each one of the data sources relevant to answer Q.

Next consider:

T/@*ENTITY*: value of the *ENTITY* attribute of the *SELECT* element *T*

T/ATTRIBUTE: value of an *ATTRIBUTE* element of the *SELECT* element *T*

T/WHERE: value of a *WHERE* element of the *SELECT* element *T*

A/text( ): retrieves the content of the *ATTRIBUTE* element *A*

ECA(S_i, E_m): retrieves the entity correspondence assertion $C\ E_m \cong E_j$, where $E_j \in S_i.E$ ($S_i.E$ is the set of entity types of the source achema S_i)

ACA(S_i, A_m): retrieves the attribute correspondence assertion $C\ A_m \cong A_j$, where $A_j \in E'.A$ and $E' \in S_i.E$ ($E.A$ is the set of attributes of the entity type E')

$X_1 \square X_2$: concatenation operator that joins two XEQ expressions X_1 and X_2

During the third phase of the query reformulation process the **ReformulateQuery** algorithm is executed for each one of the relevant data sources $\{DS_1,\ldots, DS_n\}$. The first task of this algorithm consists in reformulating the element (*X/Exp/SELECT*) that defines the mediation entity E_m being queried. To do this, the algorithm calls the procedure **ReformulateSelect**. This procedure recursively analyses the nested *SELECT* elements of T (T = *X/Exp/SELECT*) in order to identify the subentities $\{E_1,\ldots, E_n\}$ and attributes $\{A_1,\ldots, A_k\}$ of the mediation entity E_m such that *T/@entity* = E_m. For each mediation entity E_j the algorithm **ReformulateSelect** obtains the correspondence assertion which specifies how to compute E_i over the data source DS_i. In the same way for each attribute A_j, the algorithm obtains the correspondence assertion which specifies how to compute A_j over the data source DS_i. The algorithm also analyses *WHERE* elements to transform the constraints applied on the mediation elements into constraints on the source elements. This transformation is done using the algorithm **ReformulateWhere**. XEQ elements of the resulting source expression are created by the algorithm **AddPathToQuery**. The **ReformulateQuery** and **ReformulateSelect** algorithms are presented in Figure 3.6. More information about the other algorithms may be found in [Costa 2005].

```
Algorithm ReformulateQuery(XQ, Si) {Si is the X-Entity schema of the
                                    data source DSi}
 T := XQ/Exp/SELECT  {Exp is root element of a XEQ expression}
 LSi := new XEQ expression
 P := empty path
 ReformulateSelect(Lsi, T, Si, P)
 Return LSi
End

Algorithm ReformulateSelect(XSi, T, Si, P) {XSi is the XEQ expression
                                    corresponding to the data source Si }
 P:= P + T/@ENTITY
 NS:= AddPathtoQuery(XSi, ECA(Si, P))
 For each A = T/ATTRIBUTE do
    AddPathtoQuery(EXPLS, ACA(Si, P + A/text()))
 For each T' = T/SELECT do ReformulateSelect(XSi, T', Si, P)
 If (W = T/WHERE) ≠ ∅ then  XSi:= XSi ΦReformulateWhere(NS, W, Si, P)
End
```

Figure 3.6 - Query reformulation Algorithms

3.5.1 Summary

In our approach, the user query may be specified in any high level query language. This query is translated to XEQ and it can be easily rewritten into a set of local subqueries. As XEQ is an XML-based language we can take advantage of XML benefits as the capability to easily transform XML data into different formats and to easily navigate through its hierarchical structure. This work was very important to allow the generation of the first version of *Integra* prototype. It was part of the MSc thesis of Thiago Costa [Costa 2005, Loscio et al. 2006].

3.6 Quality of the Integrated Schema

Information quality (IQ) has become a critical aspect in organizations and, consequently, in Information Systems research. The notion of IQ has only emerged during the past ten years and shows a steadily increasing interest. IQ is a multidimensional aspect and it is based in a set of dimensions or criteria. The role of each one is to assess and measure a specific IQ aspect [Wang & Strong 1996, Tayi & Ballou 1998, Ballou & Pazer 1985].

In data integration systems, Naumann and Leser [Naumann, & Leser 1999] define a framework addressing the IQ of query processing. This approach proposes the interleaving of query planning with quality considerations and creates a classification with twenty two dimensions divided into three classes: one related to the user preferences, the second class concerns the query processing aspects and the last one is related to the data sources. Other relevant topic to consider is the set of quality criteria for schemas. IQ aspects of schema equivalence and transformations exploit the use of normalization rules to improve IQ in conceptual database schemas [Assenova, & Johanneson 1996]. The work proposed by Herden [Herden 2001] deals with measuring the quality of conceptual database schemas. In this approach, given a quality criterion, the schema is reviewed by a specialist in the mentioned criterion. In [Si-Said & Prat 2003] the authors propose IQ evaluation for data warehouse schemas focusing on the analyzability and simplicity criteria. Peralta et al. [Peralta et al. 2004] propose addressing the problem of data quality evaluation by a framework which is based on a graph model of the data integration system. The system is modeled as a workflow represented by a graph in which the activities perform the different tasks that extract, transform and convey data to users. It was presented an experiment with the data freshness IQ criteria. The work described in [Marotta & Ruggia 2005] also uses the activities graph representation for the integration system defined in [Peralta et al. 2004]. The authors defined the actual values of the quality properties at the sources, and at the integrated system there are the expected values of these properties.

Our contribution is the proposal of IQ criteria analysis in a data integration system, mainly related to the system schemas. Data integration systems may suffer with lack of quality in produced query results. They can be outdated, erroneous, incomplete, inconsistent, redundant, and so on. As a consequence, the query execution can become rather inefficient. To minimize the impact of these problems, we propose a quality approach that serves to analyze and improve the integrated schema definition and consequently, the query execution. Our hypothesis was that an acceptable alternative to optimize query execution would be the construction of good schemas, with high quality scores. We focused on the formal specification of algorithms and definitions of three schema IQ criteria: *schema completeness, minimality* and *type consistency* [Batista & Salgado 2007a, Batista & Salgado 2007b, Batista & Salgado 2007c]. Table 3.2 lists each criterion with its definition and the metric used to calculate scores.

Table 3.2 - IQ Criteria for schemas quality analysis

IQ Criteria	Definition	Metrics
Schema Completeness	The extent to which entities and attributes of the application domain are represented in the schema	1 – (#incomplete items / #total items)
Minimality	The extent in which the schema is modeled without redundancies	1 – (#redundant schema elements/# total schema elements)
Type Consistency	Data type uniformity across the schemas	1 – (#inconsistent schema elements / #total schema elements)

# denotes the expression "*Number of*"

Schema Completeness
The completeness can be measured as the percentage of real-world objects modeled in the integrated schema that can be found in the sources. Therefore, the schema completeness criterion is the number of concepts provided by the schema with respect to the application domain.

Minimality
Minimality is the extent in which the schema is compactly modeled and without redundancies. In our point of view, the minimality concept is very important to data integration systems because the integrated schema generated by the system may have redundancies. The key motivation for analyzing minimality is the statement that the more minimal the integrated schema is, the least redundancies it contains, and, consequently, the more efficient the query execution becomes [Kesh 1995]. Our hypothesis is that the minimality analysis will help decreasing the extra time spent by mediator with access to unnecessary information represented by redundant schema elements.

Type Consistency
Type consistency is the extent in which the attributes corresponding to the same real world concept are represented with the same data type across all schemas of a data integration system.

In what follows we will only concentrate on details of the specification and implementation of the minimality criterion, the one which presented more significant results.

3.6.1 Miminality

We have based our analysis in the measurement of minimality to help decreasing the extra time spent by mediator with access to unnecessary information represented by redundant schema elements. The minimality IQ may be useful in any integrated schema to minimize problems resulting from schema integration processes, for example, to have semantically equivalent concepts represented more than once in one schema.

We have considered an existent data integration system, formally defined as follows:

Definition 1 – Data Integration System (Ð):
A data integration system is a tuple, $\mathbf{Ð = <\delta, S_m>}$ where: $\boldsymbol{\delta}$ is the set of S_i data sources schemas, i.e. $\boldsymbol{\delta = <S_1, S_2, \ldots, S_w>}$, where w is the number of data sources in **Ð** and $\mathbf{S_m}$ is the integrated schema, generated by modules of **Ð**. In **Ð**, the following statements are true:

- S_m is a X-Entity integrated schema such as $S_m = <E_1, E_2, \ldots, E_{n_m}>$ where E_k is a mediation entity $(1 \leq k \leq n_m)$, and n_m is the number of entities in S_m;
- $\forall E_k \in S_m$, $E_k(\{A_{k1}, A_{k2}, \ldots, A_{ka_k}\}, \{R_{k1}, R_{k2}, \ldots, R_{kr_k}\})$, where $\{A_{k1}, A_{k2}, \ldots, A_{ka_k}\}$ is the set of attributes of E_k, $(a_k > 0)$; $\{R_{k1}, R_{k2}, \ldots, R_{kr_k}\}$ is the set of relationships of E_k, $(r_k \geq 0)$.
- If X_1 and X_2 are schema elements (attributes, relationships or entities), the schema mapping $X_1 \equiv X_2$ specifies that X_1 and X_2 are *semantically equivalent*, i.e., they describe the same real world concept and have the same semantics.

In data integration context, we define a schema as *redundant* if it has occurrences of redundant entities and/or relationships. We introduce the definitions 2 to 5.

Definition 2 – Redundant attribute in a single entity:

An attribute A_{ki} of entity E_k, is *redundant*, i.e., $\mathbf{Red(A_{ki}, E_k)=1}$, if $\exists E_k.A_{kj}$, $j \neq i$, $A_{kj} \in \{A_{k1}, A_{k2}, \ldots, A_{ka_k}\}$ such as $E_k.A_{ki} \equiv E_k.A_{kj}$, $1 \leq i, j \leq a_k$

Definition 3 – Redundant attribute in different entities:

An attribute A_{ki} of the entity E_k, $A_{ki} \in \{A_{k1}, A_{k2}, \ldots, A_{ka_k}\}$ is redundant, i.e. $\mathbf{Red(A_{ki}, E_k) = 1}$, if: $\exists E_o$, $o \neq k$, $E_o \in S_m$, $E_k \equiv E_o$, $E_o(\{B_{o1}, B_{o2}, \ldots, B_{oa_o}\})$, B_{oj} are attributes of E_o and $\exists E_o.B_{oj}$, $B_{oj} \in \{B_{o1}, B_{o2}, \ldots, B_{oa_o}\}$ such as $E_k.A_{ki} \equiv E_o.B_{oj}$, $1 \leq i \leq a_k$, $1 \leq j \leq a_o$. If for an attribute A_{ki} of entity E_k, $Red(A_{ki}, E_k) = 0$, we say that A_{ki} is *non-redundant*.

Definition 4 – Entity Redundancy Degree:

An entity E_k has a positive redundancy degree in schema S_m, i.e. $Red(E_k, S_m) > 0$, if E_k has at least one redundant attribute. The redundancy degree is calculated by the following formula:

$$Red(E_k, S_m) = \frac{\sum_{i=1}^{a_k} Red(A_{ki}, E_k)}{a_k}, \quad (1)$$

where

$\sum_{i=1}^{a_k} Red(A_{ki}, E_k)$ is the number of redundant attributes in E_k and;

a_k is the total number of attributes in E_k.

Definition 5 – Redundant Relationship:

Consider a relationship $R \in S_m$ between the entities E_k and E_y represented by the path $E_k.R.E_y$, $R \in \{R_{k1}, \ldots, R_{kr_k}\}$ and $R \in \{T_{y1}, \ldots, T_{yr_y}\}$, where $\{R_{k1}, \ldots, R_{kr_k}\}$ is the set of relationships of E_k and $\{T_{y1}, \ldots, T_{yr_y}\}$ is the set of relationships of E_y.

The relationship R connects E_k and E_y if and only if $R \in \{R_{k1}, \ldots, R_{kr_k}\}$ and $R \in \{T_{y1}, \ldots, T_{yr_y}\}$.

We define R as a *redundant relationship* in S_m, i.e. $Red(R, S_m) = 1$ if:
$\exists P_1$, $P_1 = E_k.R_j.\ldots.T_s.E_y$, P_1 is a path with $R_j \in \{R_{k1}, \ldots, R_{kr_k}\}$ and $T_s \in \{T_{y1}, \ldots, T_{yr_y}\}$, such that $P_1 \equiv R$.

In other words, a relationship between two entities is redundant if there are other semantically equivalent relationships which paths are connecting the same two entities. It is important to say that an equivalence relationship is determined by a path equivalence, i.e., two relationships are semantically equivalents if their paths are also semantically equivalent.

A schema is minimal if all of the domain concepts relevant for the application are described only once. Thus, we can say that the minimality of a schema is the degree of absence of redundant elements in the schema. To measure the minimality, we must first determine the redundancy degree of the schema. To each one of the next redundancy definitions (6 and 7), we assume the following:

i) n_{rel} is the total number of relationships in S_m;

ii) n_m is the total number of entities in S_m;

iii) r_k is the number of relationships of each entity E_k in S_m.

Definition 6 – Entity Redundancy of a Schema:

The total entity redundancy of a schema S_m is computed by the formula:

$$ER(S_m) = \frac{\sum_{k=1}^{n_m} Red(E_k, S_m)}{n_m} \quad (2),$$

where $Red(E_k, S_m)$ is the redundancy degree of each E_k in S_m.

Definition 7 – Relationship Redundancy of a Schema:

The relationship redundancy degree of S_m is measured by the equation:

$$RR(S_m) = \frac{\#\,Red(R, S_m)}{n_{rel}} \quad (3),$$

where $\#Red(R, S_m)$ is the number of redundant relationships in S_m as stated in Definition 5.

Definition 8 – Schema Minimality:

We define the overall redundancy of a schema in a data integration system as the sum of the aforementioned redundancy values: entities (ER) and relationships (RR), by the formula:

$$Mi_{S_m} = 1 - [ER(S_m) + RR(S_m)] \quad (4)$$

3.6.2 Schema IQ Improvement

After detecting the schema IQ anomalies, it is possible to restructure it to achieve better IQ scores. In order to improve minimality scores, redundant elements must be removed from the schema. We proposed schema improvement actions specified in the algorithm of Table 3.3.

Table 3.3 - Schema improvement algorithm

1	Calculate minimality score and if minimality = 1, then stop;
2	Search for fully redundant entities in S_m;
3	If there are fully redundant entities then eliminate the redundant entities from S_m;
4	Search for redundant relationships in S_m;
5	If there are redundant relationships then eliminate the redundant relationships from S_m;
6	Search for redundant attributes in S_m;
7	If there are redundant attributes then eliminate the redundant attributes from S_m;
8	Go to Step 1

The condition of minimality = 1 is the ideal case where the schema is minimal, and this can occur when all schema redundancies are eliminated. The detection of redundant elements processes are executed in steps 2, 4 and 6, already described in previous definitions. Redundancies elimination in steps 3, 5 and 7 are discussed in the following.

Redundant Entities Elimination

After removing a redundant entity E, its relationships must be relocated to a semantic equivalent remaining entity. When removing a redundant entity E_1 ($E_1 \equiv E_2$), the *IQ Manager* transfers the relationships of E_1 to the remaining equivalent entity E_2. Three different situations may occur when moving a relationship R_x, $R_x \in E_1$:

i) if $R_x \in E_2$ then R_x is deleted because it is no longer necessary;

ii) if $R_x \notin E_2$ but $\exists R_y$, $R_y \in E_2$ such as $R_x \equiv R_y$ then R_x is deleted;

iii) if $R_x \notin E_2$ and there is no R_y, $R_y \in E_2$ such as $R_x \equiv R_y$, then R_x is connected to E_2.

The first and second situations are not supposed to cause any other schema modification besides the entity deletion. The third case needs more attention, once redundant relationships of the removed entity have to be relocated as stated as follows.

Definition 9 – Substitute Entity:

E_k is a fully redundant entity, if and only if $Red(E_k, S_m)=1$ and E_k has at least one *Substitute Entity* E_s, i.e. $Subst(E_k) = E_s$, such as:

- $E_k(\{A_{k1}, \ldots, A_{ka_k}\}, \{R_{k1}, \ldots, R_{kr_k}\})$ A_{kx} are attributes and R_{ky} are relationships of E_k and;
- $E_s(\{A_{s1}, \ldots, A_{sa_s}\}, \{R_{s1}, \ldots, R_{sr_s}\})$ A_{sz} are attributes and R_{st} are relationships of E_s and
- $E_k \equiv E_s$ and $\forall E_k.A_{ki} \in \{A_{k1}, \ldots, A_{ka_k}\}$, $\exists E_s.A_{sj} \in \{A_{s1}, \ldots, A_{sa_s}\}$ with $E_k.A_{ki} \equiv E_s.A_{sj}$.

An entity E_k is considered fully redundant when all of its attributes are redundant, i.e. $\mathbf{Red(E_k, S_m)=1}$ and it has a substitute entity E_s in S_m. All the attributes of E_k are contained in E_s. E_k may be removed from the original schema S_m without lost of relevant information if it is replaced by its *substitute entity* E_s. Any existing relationship from E_k may be associated to E_s.

Definition 10 – Relationship Relocation:

In a schema S_m, if $Subst(E_k)=E_s$, then E_k can be eliminated from S_m. In this case, in order to do not lose any information, E_k relationships may be relocated in S_m. It is possible to relocate the relationships from E_k to E_s according to the following rules, i.e. $\forall E_k.R_{kj}$:

i. If $E_k.R_{kj} \in \{R_{s1}, \ldots, R_{sr_s}\}$ then R_{kj} must be deleted because it is no longer useful;
ii. If $E_k.R_{kj} \notin \{R_{s1}, \ldots, R_{sr_s}\}$ but $\exists E_s.R_{sp}$, such that $E_k.R_{kj} \equiv E_s.R_{sp}$ then $E_k.R_{kj}$ must be deleted because it has an equivalent relationship in E_s;
iii. If $E_k.R_{kj} \notin \{R_{s1}, \ldots, R_{sr_s}\}$ and $\nexists E_s.R_{sp}$ such as $E_k.R_{kj} \equiv E_s.R_{sp}$ then, E_s is redefined as $E_s(\{A_{s1}, \ldots, A_{sa_s}\}, \{R'_{s1}, \ldots, R'_{sr_s}\})$, A_{sz} are attributes and R'_{st} are relationships of E_s and $\{R'_{s1}, \ldots, R'_{sr_s}\} = \{R_{s1}, \ldots, R_{sr_s}\} \cup R_{kj}$.

The relationship relocation is illustrated in Figure 3.7.

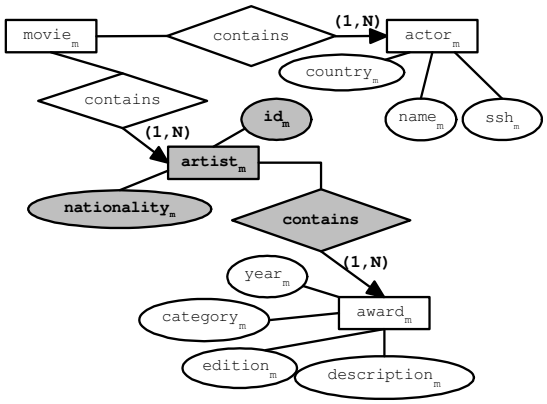


Figure 3.7. Redundant entity elimination

The fully redundant entity $artist_m$ (with its attributes) is removed and it is substituted by the semantically equivalent $actor_m$. Consequently, the relationship $movie_m\_artist_m$ may be deleted replaced by the remaining equivalent relationship $movie_m\_actor_m$. The relationship $artist_m\_award_m$ is relocated to $actor_m$, turning into the new relationship $actor_m\_award_m$. With these operations, it is possible to obtain a no redundant schema.

Redundant Relationships Elimination

After removing redundant entities and performing the necessary relationship relocations, the *IQ Manager* is supposed to analyze if there are remaining redundant relationships to eliminate them. This can be accomplished by purely deleting from the schema, the relationships identified as redundant. After eliminating the redundant relationships the schema becomes with no relationship redundancies and do not have had lost of relevant information.

Redundant Attributes Elimination

The last step of schema improvement algorithm consists in investigating and eliminating remaining redundant attributes in schema. Similarly to the redundant relationships removal step, these attributes may merely be deleted from schema. This occurs because the schema always has semantically equivalent attributes to substitute the redundant ones. After executing the schema improvement steps, the *IQ Manager* can recalculate and analyze minimality scores in order to determine if the desired IQ is accomplished.

3.6.3 Summary

The quality analysis is performed by a software module called *IQ Manager* or *Information Quality Manager* which may be attached to a data integration system. At the moment of integrated schema generation or update, this module proceeds with the criteria assessment and then, according to the obtained IQ scores, may execute adjustments over the schema to improve its design and, consequently, the query execution. The specification and evaluation of the minimality criterion, along with the other two defined criteria (type consistence and completeness), are part of the ongoing PhD thesis of Maria da Conceição Batista [Batista 2008, Batista & Salgado 2007a, Batista & Salgado 2007b, Batista & Salgado 2007c].

3.7 Main Results

There are several results related to the research done about data integration systems. In addition to an Integra prototype that allowed the experimentation and validation of our proposal, some PhD and MSc theses were concluded. The main results were published in international and Brazilian conferences. In what follows we present each one of these results.

3.7.1 Integra Prototype

An *Integra* prototype was implemented allowing the execution of the whole query process: from the submission of a user query to the answer of the integrated results. Wrappers were implemented to translate queries from XEQ to the data sources' native language. In the initial version of the prototype, only wrappers for relational databases were implemented. During the tests, we used two real databases that stores medical information: Healthnet[2] (data of a public hospital) and TELEMED[3] (data obtained from video-conferencing sessions of real-time consultations between medical specialists in different locations), both implemented in MySQL. In the current version of the prototype, the mediation schema and the correspondences between the mediation schema and the source schemas are manually defined and stored in the *Mediator Knowledge Base*. The queries were translated to SQL using a XSLT stylesheet, and it is submitted to the data source through a JDBC driver.

[2] http://www.nutes.ufpe.br/servicos/health.html

[3]Database of a Brazilian hospital (Hospital Português – Recife)

The IQ Manager module was integrated to the currently version of *Integra*. It was written in Java and the experiment used two databases – MySQL and PostgreSQL – to store the data sources. For minimality improvement, the experiment was done in the following steps: (i) initially, the queries were submitted over an integrated schema 26% redundant and the execution times were measured; (ii) the redundancy elimination algorithm was executed over the redundant integrated schema generating a minimal schema (100% of minimality); (iii) the same queries of step (i) were re-executed. The results obtained with these experiments, using the two real world data sources described above, have been satisfactory since query performance was improved. In our experimentation, we choose four types of user queries and executed the same query over redundant schemas. Each query was submitted five times, and its processing times were computed. In all of the four queries (simple selection, selection with condition, join with one condition and join with two conditions) the average execution time was lower when the integrated schema is minimal. In all of the tested cases, the results confirmed the existence of improvements in query execution time. The four query execution times are summarized in Table 6 and in Figure 3.8. By comparing the results, it is possible to see that the query performance was improved in an average time of **45,40%.**

Table 3.4 - Summary of query execution times

Query		Average Times (Sec)		Performance
		Redundant Schema	Minimal Schema	Gain(%)
UQ1	Simple selection	69,3720	37,0094	46,65%
UQ2	Selection with condition	11,8312	6,8970	41,70%
UQ3	Join with two conditions	35,9128	13,5904	62,16%
UQ4	Join with one condition	12,0776	8,3248	31,07%
		Average Gain		**45,40%**

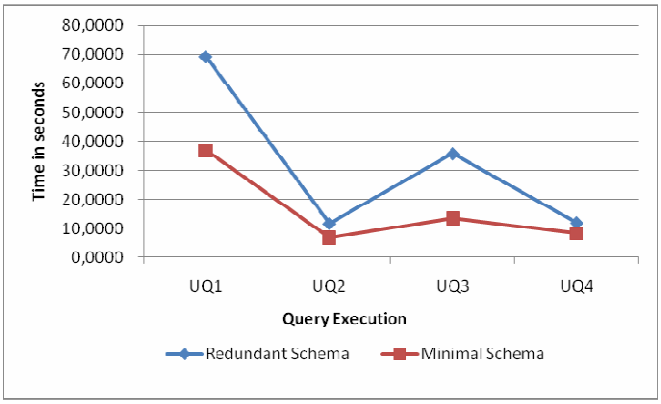


Figure 3.8 - Summary of execution times of UQ

3.7.2 PhD and MSc Theses

The researches in data integration issues were done along with two PhD students (one is finishing in a few months) and five MSc students as listed in chronological order in the following. It is important to notice that the implementation results were accomplished due to other undergraduate students not nominated in this list.

[Loscio 2003] Bernadette Farias Lóscio. Managing the Evolution of XML-based Mediation Queries, PhD, 2003

[Batista 2003] Maria da Conceição Moraes Batista. Otimização de Acesso em um Sistema de Integração de Dados através do Uso de *Caching* e Materialização de Dados, MSc, 2003

[Cardoso 2004] Rafael Cunha Cardoso. Um Sistema de Recuperação e Extração de Informação Utilizando Conceitos da Web Semântica, MSc, 2004

[Costa 2005] Thiago Alves Costa. Gerenciamento de Consultas em um Sistema de Integração de Dados, MSc, 2005

[Amaral 2007] Haroldo José Costa do Amaral. Materialização Seletiva de Dados Baseada em Critérios de Qualidade, MSc, 2007

[Galvao 2007] Walter de Carvalho Mattos Galvão. Uma proposta pra o Gerenciamento de Cache de um Sistema de Integração de Dados, MSc, 2007

[Batista 2008] Maria da Conceição Moraes Batista. Schema Data Quality on Information Integration Systems, PhD, 2004-2008 (ongoing)

3.7.3 Publications

Among the thirteen papers listed below, we have one journal paper at *Journal of the Brazilian Computer Society* (internationally indexed) and two submitted to *ACM Journal on Data and Information Quality* and to *Journal of Information Assurance and Security*. In addition, we have four papers at known international conferences: *CoopIS*, *ICEIS*, *ACM SAC* and *IEEE ICDIM*, others at international workshops associated to conferences and at the *Brazilian Symposium on Databases*.

[Batista & Salgado 2008a] Batista, M. C. M., Salgado, A. C. Data Integration Schema Analysis: An Approach with Information Quality, *ACM Journal on Data and Information Quality*, edited by ACM, 2008 (submitted to evaluation)

[Batista & Salgado 2008b] Minimality Quality Criterion Evaluation for Integrated Schemas, *Journal of Information Assurance and Security*, edited by Dynamic Publishers Inc., 2008 (submitted to evaluation)

[Batista & Salgado 2007a] Batista, M. C. M., Salgado, A. C. Data Integration Schema Analysis: An Approach with Information Quality, In Proc of the 12th *International Conference on Information Quality*, MIT, Cambridge, USA, 2007. p.1 – 8

[Batista & Salgado 2007b] Batista, M. C. M., Salgado, A. C. Information Quality Measurement in Data Integration Schemas In Proc of the *5th International Workshop on Quality in Databases*, Vienna, Austria, 2007. p.61 – 72

[Batista & Salgado 2007c] Batista, M. C. M., Salgado, A. C., Minimality Quality Criterion Evaluation for Integrated Schemas, In Proc. of the *2nd IEEE International Conference on Digital Information Management*, Lyon, France, 2007, p.436 – 441

[Loscio et al. 2006] Loscio, B. F., Costa, T. A., Salgado, A. C., Freitas, J. S. Query Reformulation for an XML-based Data Integration System In: The *21st Annual ACM Symposium on Applied Computing*, Dijon, France, 2006, p.498 – 502

[Loscio & Salgado 2004] Loscio, B. F., Salgado, A. C., Evolution of XML-based Mediation Queries in a Data Integration System, In Proc. of the *3rd International Workshop on Evolution and Change in Data Management*, . LNCS 3289, Shanghai, China, 2004, p.402 – 414

[Loscio et al. 2003] Loscio, B. F., Galvao, L. R., Salgado, A. C. Conceptual Modeling of XML Schemas, In proc. of the *5th International Workshop on Web Information and Data Management (WIDM'03)*, New Orleans, USA, 2003. p.102 - 105

[Loscio & Salgado 2003] Loscio, B. F., Salgado, A. C. Generating Mediation Queries for XML-based Data Integration Systems In Proc. of the 18th *Brazilian Symposium on Databases (SBBD)*, Manaus, Brazil, 2003. p.99 – 113

[Bouzeghoub et al. 2003] Bouzeghoub, M., Loscio, B. F., Kedad, Z., Salgado, A. C. Managing the Evolution of Mediation Queries In Proc. of the *10th International Conference on Cooperative Information Systems*, LNCS 2888, Catania, Italy, 2003, p.22 – 37

[Batista el al. 2003] Batista, M. C. M., Loscio, B. F., Salgado, A. C. Optimizing Access in a Data Integration System with Caching and Materialized Data, In Proc. of the *5th International Conference on Enterprise Information Systems*, Anvers, France, 2003. p.529 – 532

[Batista & Salgado 2003] Batista, M. C. M., Salgado, A. C. Using Quality Criteria to Selective Data Materialization, In Proc. of the *18th Brazilian Symposium on Databases (SBBD)*, Manaus, Brazil, 2003, p.72 – 83

[Loscio et al. 2002] Loscio, B. F., Salgado, A. C., Vidal, V. M. P. Using Agents for Generation and Maintenance of Mediators. *Journal of the Brazilian Computer Society.* , v.8, 2002 p.32 - 42

[Loscio et al. 2001] Loscio, B. F., Salgado, A. C., Vidal, V. M. P. Using Agents for Generation and Maintenance of Mediators in a Data Integration System In Proc. of the *16th Brazilian Symposium on Databases (SBBD)*, , Rio de Janeiro, Brazil 2001. p.172 – 186

[Vidal et al.] Vidal, V. M. P., Loscio, B. F., Salgado, A. C. Using Correspondence Assertions for Specifying the Semantics of XML-based Mediators In Proc. of the *International Workshop on Information Integration on the Web*, Rio de Janeiro, Brazil, 2001. p.3 - 11

3.8 References

[Abiteboul et al. 1999] Abiteboul, S., Buneman, P. and Suciu, D. *Data on the Web: From Relations to Semistructured Data and XML.* 1st Edition. Morgan Kaufmann Publishers, 1999.

[Ambite et al. 2001] Ambite, J., Knoblock, C., Muslea, I., Philpot. A.: Compiling Source Description for Efficient and Flexible Information Integration, *Journal of Intelligent Information Systems*, Vol. 16, N. 2, 2001, p.149-187

[Assenova, & Johanneson 1996] Assenova, P., Johanneson, P. Improving Quality in Conceptual Modeling by the Use of Schema Transformations, In Proc. of the *15th Int. Conf. of Conceptual Modeling (ER´96)*, Cottbus, Germany, 1996, p.277-291

[Ballou & Pazer 1985] Ballou, D.P., Pazer, H.L. Modeling Data and Process Quality in Multi-input, Multi-output Information Systems, *Management Science,* Vol.31, N.4, 1985, p.150-162.

[Bergamaschi et al. 1998] Bergamaschi, S., Castano, S., De Capitani Di Vimercati, S., Montanari, S., Vincini, M. A Semantic Approach to Information Integration: the MOMIS Project, In Proc. of the *Sesto Convegno della Associazione Italiana per l'Intelligenza Artificiale*, Padova, Italy, 1998

[Boyd et al. 2004] Boyd, M., Kittivoravitkul, S., Lazanitis, C., McBrien, P., Rizopoulos, N. AutoMed: A BAV Data Integration System for Heterogeneous Data Sources. In Proc. of the *16th Advanced Information Systems Engineering Conference (CAiSE),* Riga, Latvia, 2004, p.82-97.

[Bray et al. 2006] Bray, T., Paoli, J., Sperberg-McQueen, C.M, Maler, E., Yergeau, F. Extensible Markup Language (XML) 1.0 (Fourth Ed.), *W3C*, 2006 (available on http://www.w3.org/TR/REC-xml, last access February 2008)

[Chawathe et al. 1994] Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J. and Widom, J. The TSIMMIS Project: Integration of Heterogeneous Information Sources. In Proc. of the *10th Meeting of the Information Processing Society of Japan* (IPSJ), Tokio, Japan, 1994, p.7-18.

[Chen 1976] Chen, P.P. The Entity-Relationship Model: Toward a unified view of data, *ACM Transactions on Database Systems*, Vol. 1, N. 1, 1976, p.1-36

[Draper et al. 2001] Draper, D., Halevy, A. Y. and Weld D. S.. The Nimble XML Data Integration System^{tm}. In Proc. of the *17th International Conference. on Data Engineering* (ICDE), Heildelgerg, Germany, 2001, p. 155-160.

[Deutsch & Tannen 2003] Deutsch, A., Tannen, V. Reformulation of XML Queries and Constraints. In Proc. of the *9th International Conference on Database Theory (ICDT),* Siena, Italy, 2003, p.255–241.

[Fallside & Walmsley 2004] Fallside, D., C., Walmsley, P. XML Schema Part 0: Primer Second Edition. Available at: http://www.w3.org/TR/xmlschema-0/, 2004

[Friedman et al. 1999] Friedman, M., Halevy, A. Y., Millstein, T. D.: Navigational plans for data integration. In Proc. of *Workshop on Intelligent Information Integration (IJCAI-99)*, Stockholm, Sweden, 1999

[Gardarin et al. 2002] Gardarin, G., Mensch, A., Tomasic, A., An Introduction to the e-XML Data Integration Suite, In Proc. of *8th International Conference on Extending Database Technology (EDBT)*, Prague, Czech Republic, 2002, p. 297-306

[Halevy 2000] Halevy, A., Y.: Logic-based Techniques in Data Integration. J. Minker, editor *Logic based Artificial Intelligence*, Kluwer Publishers, 2000

[Hammer et al. 1997] Hammer J., Brennig, M., Garcia-Molina, H., Nesterov, S., Vassalos, V., Yerneni, R. Template based wrappers in the tsimmis system, In Proc. of the *ACM SIGMOD International Conference on Management of Data*, Tucson, USA, 1997, p.532-535.

[Herden 2001] Herden, O. Measuring Quality of Database Schema by Reviewing - Concept, Criteria and Tool, In Proc. of the *5th Intl Workshop on Quantitative Approaches in Object-Oriented Software Engineering*, Budapest, Ungarn, 2001, p.50-79.

[Kedad & Bouzeghoub 1999] Kedad, Z., Bouzeghoub, M., Discovering View Expressions from a Multi-Source Information System, In Proc. of the *4th IFCIS International Conference on Cooperative Information Systems (CoopIS)*, Edinburgh, Scotland, 1999, p. 57-68

[Kesh 1995] Kesh, S. Evaluating the Quality of Entity Relationship Models. *Information and Software Technology*, Vol.37, N.2, 1995, p. 681-689

[Kotonya & Sommerville 1997] Kotonya, G., Sommerville, I. Requirements Engineering: Processes and Techniques. 1stEdition, Wiley & Sons, 1997

[Labio et al. 1997] Labio, W. J., Zhuge, Y., Wiener, J. L., Gupta, H., Garcia-Molina, H. and Widom, J. The WHIPS Prototype for Data Warehouse Creation and Maintenance. In Proc. of the *ACM SIGMOD International Conference on Management of Data*, Tucson, USA, 1997, p.557-559.

[Lenzerini 2002] Lenzerini, M.: Data Integration: A Theoretical perspective. In Proc. of the *21th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 2002)*, Madison, USA, 2002, p.233-246

[Marotta & Ruggia 2005] Marotta, A. And Ruggia, R. Managing Source Quality Changes in Data Integration Systems. In Proc. of the 2nd *International Workshop on Data and Information Quality (DIQ'05)*, Porto, Portugal, 2005

[McBrien & Poulovassilis 2002] McBrien, P., Poulovassilis, A.: Schema Evolution in Heterogeneous Database Architectures, A Schema Transformation Approach, In Proc. of the

Conference on Advanced Information Systems Engineering (CaiSE 2002), Toronto, Canada, 2002, p.484-499

[McBrien & Poulovassilis 2003] McBrien, P., Poulovassilis, A.: Data integration by Bi-directional Schema Transformation Rules, In Proc. of the *19th International Conference on Data Engineering, (ICDE),* Bangalore, India, 2003, p.227-238.

[Naumann, & Leser 1999] Naumann, F., Leser, U. Quality-driven Integration of Heterogeneous Information Systems, In Proc. of the *25th International Conference on Very Large Data Bases (VLDB)*, Edinburgh, Scotland, 1999, p.447-458.

[Nica & Rundensteiner 1999] Nica, A., Rundensteiner, E. A.: View maintenance after view synchronization, In Proc. of the *International Database Engineering and Application Symposium (IDEAS'99)*, Montreal, Canada, 1999, p.215-213

[Peralta et al. 2004] Peralta, V., Ruggia, R., Kedad, Z., Bouzeghoub, M. A Framework for Data Quality Evaluation in a Data Integration System, In Proc. of the *19th Brazilian Symposium on Databases (SBBD'2004)*, Brasilia, Brazil, 2004, p.121-133.

[Shanmugasundaram et al. 2001] Shanmugasundaram, J, Kiernan, J., Shekita, E.J., Fan C., Funderburk J. Querying XML Views of Relational Data. In Proc. of the *27th International Conference on Very Large Data Bases (VLDB)*, Roma, Italy, 2001, p.261-270

[Si-Said & Prat 2003] Si-Said, S. C., Prat, N. Multidimensional Schemas Quality: Assessing and Balancing Analyzability and Simplicity, In Proc. of the *International Workshop on Conceptual Modeling Quality -ER'03*, LNCS 2814, 2003, p.140-151

[Tayi & Ballou 1998] Tayi, G. K., Ballou, D. P. Examining Data Quality. *Communications of the ACM*, Vol. 41, N. 2, 1998, p.54-57

[Ullman 1997] Ullman, J. D.: Information integration using logical views. In Proc. of the *6th International Conference on Database Theory (ICDT'97)*, Delphi, Greece, 1997, p.19-40

[Wang & Strong 1996] Wang R.Y., Strong D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, Vol.12, N.4, 1996, p.5-33.

[Widom 1995] Widom, J. Research Problems in Data Warehouse, In: Proc of 4th Int. *Conference on Information and knowledge Management (CIKM)*, Baltimore, 1995, p. 25-30.

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

CHAPTER 4

Semantic Issues in PDMS

4.1 Introduction

Today's continuous growth of large Web applications entails an increasing need for integrating and sharing large amounts of data, which come from a number of heterogeneous, dynamic and distributed data sources. As a result, data integration solutions must deal with different types of heterogeneity: *structural*, *syntactical*, *systemic* and *semantic* [Wache et al. 2001]. However, to reconcile heterogeneity, more semantic information about data sources and other elements involved in data integration (e.g. users and applications) is needed [Kashyap & Sheth 1996].

Schemas and instances drawn from data sources rarely contain explicit semantic descriptions which could be used to derive the meaning or purpose of schema elements (e.g. entity, attribute and relationship). Implicit semantic information needs to be extracted in order to clarify the meaning of the schema elements. To achieve this, an ontology belonging to a given knowledge domain will provide the information regarding semantic relations among the vocabulary terms shared by the data sources. A domain ontology establishes a common vocabulary for information sharing in a domain including machine-interpretable definitions of concepts and relations among them [Noy & McGuiness 2001].

Semantic interpretation, however, regards people's understanding about the elements modeled, which is represented by their particular views [Ziegler & Ditrich 2004]. Furthermore, semantic interpretation is a context-dependent task which requires a specific understanding of the shared domain knowledge. *Context* may be employed as a way to improve decision-making over heterogeneity reconciliation in data integration processes since it helps to understand the data schema semantics as well as the data content semantics.

In what follows we briefly discuss about the two semantic issues we have been working on (Section 4.2). In Section 4.3 we present our proposal to a context-oriented model and a domain-independent context manager. A contextual ontology to data integration is presented in Section 4.4. The Section 4.5 is dedicated to a semantic-based approach to peers' organization in a PDMS.

We have currently a research project with financial support from a Brazilian institution (CNPq) from 2008 to 2010.

4.2 Semantic Issues

We have been working on semantic issues to improve data integration on distributed and P2P architectures. The two main concepts of our interest are ontologies and context, discussed in next subsections.

4.2.1 Ontologies

Ontology is a term well known in areas such as Philosophy and Epistemology denoting in that order, a "subject existence" and a "knowledge to know" [Chandrasekaran et al. 1999]. Recently, this term has being used in Artificial Intelligence (AI) to describe concepts and relationships used by agents. In the database community, ontology is a partial specification of a domain, which expresses entities, relationships between these entities and integrity rules.

The specification of an ontology makes possible the communication between computer systems, independently of the architecture and the information domain treated, eliminating ambiguities over a domain terminology [Bézivin 1998]. The main characteristics of an ontology are:

- *Sharing:* it certifies that the several agents interacting over a theme, possesses the same understanding of the domain concepts.
- *Filtering:* it allows the modeled domain taking into consideration just the part of reality that interest to the application, discarding many unnecessary concepts.

This way, an ontology provides a common/shared understanding about concepts of specific knowledge domains. It enables knowledge sharing between human and software agents, allows knowledge reuse between systems, and can be used by existing inference engines for reasoning.

Ontologies can also be used as a representation model. They enable the formal specification of concepts, such as entities, attributes and relationships, and ease the reuse of existing solutions.

4.2.2 Context

Context is defined as any information used to characterize the situation of an entity where an entity is a person, place, or object that is considered relevant to the interaction between a user and an application [Dey & Abowd 2000]. Context-aware systems are those able to understand the context of users and anticipate their needs, in terms of services and/or information. Moreover, context can play an important role in the communication and interaction between humans as well as between humans and machines, since it diminishes ambiguity and conflicts, increases the expressiveness of dialogues, makes applications more friendly, flexible and easy to use, and consequently raises user's satisfaction.

In our approach, *Context* is a set of elements surrounding a domain entity of interest which are considered relevant in a specific situation during some time interval. The *domain entity* of interest may be a person, a procedure, a file, a set of data or even an inter-schema mapping. Furthermore, we use the term *contextual element* (CE) referring to pieces of data, information or knowledge that can be used to define the *Context* [Vieira et al. 2007a].

Our work is based on two classical definitions of context. The first states that context is any information that can be used to characterize the situation of an entity (e.g. person, place, object, application) [Dey & Abowd 2000]. The second indicates that context is always related to a focus and that at a given focus the context is the aggregation of three types of knowledge: Contextual Knowledge (CK), External Knowledge (EK) and Proceduralized Context (PC) [Brézillon & Pomerol 1999]. An illustration of the combination of these two definitions is shown in Figure 4.1.

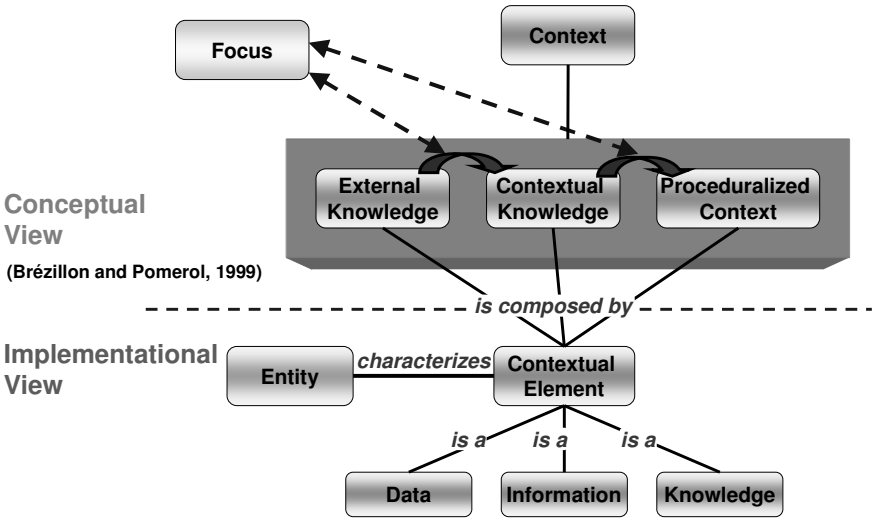


Figure 4.1 - Illustration of our definition of context

Most context-sensitive systems do not take into account requirements such as modularity, reusability or interoperability, and implement context manipulation in a proprietary way to attend to particular needs of each system. Context managers enable reuse of context-related solutions and reduce the complexity associated to building context-sensitive systems. Since CEs may come from multiple, heterogeneous sources it is important to think about formalisms and common languages that enable context sharing and interoperability of these sources to compose different systems. So, we are interested in investigating the specificities of mechanisms to manage CEs in a domain-independent way.

4.2.2 Summary

The ontologies and context concepts are being used in the specification of several approaches: a semantic name resolution process [Belian 2008], a recommender system [Petry 2007] and a cooperative learning environment [Siebra 2007]. In addition, a semantic-based approach to peer management is being proposed, including peer clustering [Pires 2009] and query reformulation [Souza 2009]. These works have motivated the definition of the domain-independent Context-Oriented Model presented in next section.

4.3 Context-Oriented Model

The Context-Oriented Model (COM) is divided into three layers (Figure 4.2): the **upper layer**, which characterizes the generic context management concepts and can be qualified as the conceptual model, since it is used for creating individual models; the **middle layer** that defines the domain related concepts in accordance with the upper layer; and the **lower layer**, which represents the concepts instantiation according to a specific application [Vieira et al. 2007b, Vieira et al. 2008, Vieira 2008].

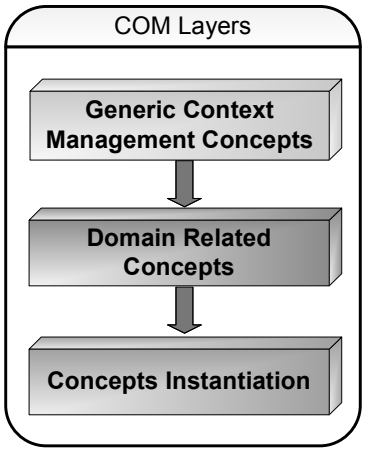


Figure 4.2 - Interaction between the three layers in the COM model

We use the following hypothesis underlying this layer division: if we define the context related concepts in a high level domain-independent layer, then these concepts can be managed in a generic manner, without worrying about the domain particularities. So, the context manager mechanisms will be applied over the upper layer concepts and a compatible context-sensitive system must be modeled by instantiating these upper concepts.

We will focus on describing the upper layer concepts, detailed in next subsections.

4.3.1. Generic Context Management Concepts Specification

An illustration of the generic context management concepts and their properties and relationships are presented in Figure 4.3. The model is centered on five **main concepts**: *Entity*, *ContextualElement*, *Focus*, *Rule* and *Action*; and three **derived concepts**: *CEFSet*, *RFSet* and *ProceduralizedContext*. The main concepts must be instantiated by the context-sensitive system designer, while the derived concepts are built by the context manager based on the main concepts and guided by the focus.

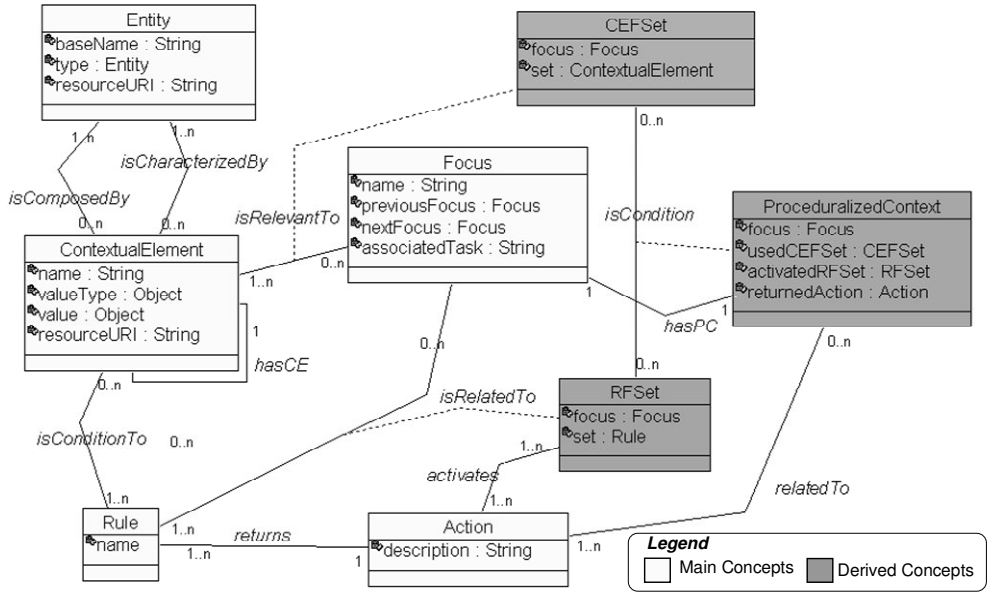


Figure 4.3 - Overview of COM model with the generic context management concepts

An entity is anything in the real world that is relevant to describe the domain, e.g. *Person*, *Hotel*, *Mission.* An Entity is defined through: a *name*; a *type* (e.g. an entity Missionary is of *type* Person); and an *URI* that allows to link the entity to an external resource that contains further knowledge, such as its OWL description file. The Entity has an association *isCharacterizedBy* with the concept *ContextualElement*, meaning that one Entity is characterized by one or more CEs.

A *ContextualElement*, as defined before, may represent a data, information or knowledge, and is used to characterize entities. The set of CEs related to an entity are composed by: (i) the attributes defined for that entity in the knowledge domain; and (ii) the relationships between that entity and other entities in the domain. For example, considering the entity Mission, its set of related CEs include: *location*, *initialDate*, *endDate*, *duration*, *whoPays*, *officialReasons*. A CE is represented by: *name*, *valueType*, *value* and *resourceURI*. The *valueType* indicates the expected value for the CE. For example, the CE *location* expects as value an entity of type Location while the CE *numOfStars* expects an object of type Integer. The attribute *value* accepts the ascribed instance of the CE and will be filled by the context-sensitive application during its usage. The attribute *resourceURI* identifies a link to an external resource that describes the CE, such as an OWL file. A CE may be associated to one or more CEs, creating a hierarchical structure between contextual elements. The CE concept has two associations with the Entity concept. The first points out that an entity *is characterized by* one or more CEs, and the second implies that a CE may be *composed by* one or more entities. The latter is important in situations where a CE associated to an Entity X needs to make a reference to an Entity Y.

The *Rule* concept was included in the model to explicitly represent the rules associated to the CEs, necessary to produce CE information from CE data and also to support the building of the Proceduralized Context in the focus. A Rule is identified by a *name*, has one or more conditions, which are represented by the association *isConditionTo* with the Contextual Element concept, and it has one returning action, represented by the association *returns* with the Action concept.

An *Action* is represented by a *description*. Actions are distinguished from rules to ease the modeling of context-dependent actions that could be implemented by the application. In this light, rules are specified having in mind the possible and desired actions.

In our model, the *Focus* is a central concept, since the context is always related to it. It is used to identify clear points of time and space that the context is all about. The focus allows the context manager to determine what CEs should be used and instantiated, since it determines the relevance of a CE in a specific situation. An example of focus is *Book Hotel*. The focus is identified by a *name*. Since it is related to objectives to solve a problem or to execute a task, we modeled it as a sequence of foci, where each element has references to the *previousFocus* and the *nextFocus*. Another attribute is the *associatedTask* that may be a problem, a decision making or a task, and can contain a textual description or an URI referencing an external resource that describes the task. The concept focus has the associations: *isRelevantTo*, with the concept *ContextualElement*, showing that the CE is relevant to that focus. Similarly the concept Rule has an association *isRelatedTo* showing that the rule is related to the focus. The *hasPC* indicates that a focus has one related *ProceduralizedContext*.

Context is a dynamic construct that evolves with the focus. As the focus changes, the set of CEs that must be considered changes accordingly [Brézillon & Pomerol 1999]. Thus the *Focus* concept is the one who guides the generation of the derived concepts (CEF-Set, RF-Set and Proceduralized Context). This is done by the methods associated to the it: *changeFocus*, shows that the focus changed pointing out that it is necessary to review the current context; *generateCEFSet*, which marks the building of the CEF-Set according to the current focus and its associated CEs; *generateRFSet*, to build the RF-Set according to the relevant Rules for the focus; and *buildPC*, which indicates the procedures to construct the PC for the focus.

4.3.2 Context Management

Context management involves the definition of models and systems to assist the acquisition, manipulation and maintenance of a shared repository of CEs, thus enabling the usage of these elements by different context-sensitive systems. The main idea is to reduce the complexity of building context-sensitive systems, by transferring tasks related to CE manipulation to an intermediate layer. In this light, the task of managing context includes the definition of: (1) a representation model to describe and share CE sets; (2) an infrastructure to detect, update and query CE sets; (3) mechanisms to process, reason about, and infer new CE sets from existing ones; and (4) mechanisms to identify the ICE in a focus.

An overview of the context management process and its main functionalities is illustrated in Figure 4.4. The first step is to *acquire* the CEs associated to a situation. Computer systems may use virtual and physical sensors, user interfaces (e.g. forms), persistent databases, and so on, to acquire these elements. After that, the system must use knowledge bases, and inference engines to *process* the acquired CEs through reasoning and associations. The interpreted context is used to infer information and to trigger services that must be provided and executed. A context management system brings the added following advantages:

- *Reusability*: the solution for each context management task can be done in a generic way and be reused by several applications;
- *Sharing*: Applications can share CEs acquired from different and heterogeneous context sources;
- *Context source independence*: Applications are developed independently from the underlying contextual source;
- *Ease of use*: Application developers can focus on their business model and leave details of context management to the manager implementation.

In order to be effective, a context manager must take into account aspects such as: separation of the context model and the application domain model; maintenance of a sharable context model that enables communication between different components or systems; provision of descriptions and interfaces for the manager internal components and their formats to allow communication and interoperability among the manager's components and context consumers.

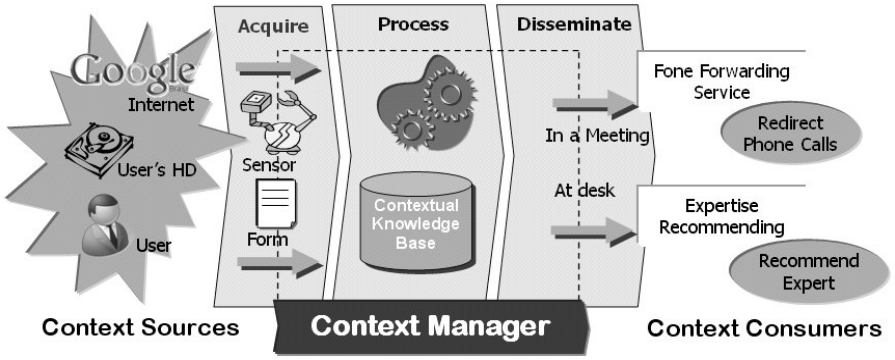


Figure 4.4 - Overview of a Generic Context Manager Main Functionalities

4.3.3 Summary

COM is an approach for context representation that proposes the separation of the context management concepts from domain and application concepts. Our hypothesis is that it is impossible to imagine a context model that is at the same time specific and generic, since context is extremely domain and application-dependent. Thus, we propose that developers of context-

sensitive systems rethink the way of modeling their systems including the context management phases in the system building process and considering the context related concepts when specifying system's functionalities. This work is part of the Vaninha Vieira's PhD thesis [Vieira 2008, Vieira et al. 2008, Vieira et al. 2007a, Vieira et al. 2007b, Vieira et al. 2005a, Vieira et al. 2005b].

4.4 Semantics in Data Integration

Recent works have considered the use of ontologies [Wache et al. 2001] as a way of providing a domain reference to improve schema and data integration. We use contextual information, i.e. the circumstantial information that makes a situation unique and comprehensible [Brézillon 2003], as well as domain ontologies, as a way to enrich the data integration process. In this light, context information is used to ease schema mapping discovery, helping to determine the correct meaning of an entity. It is also used to improve query processing capabilities, providing users with "meaningful", i.e., more relevant results. Thus, contextual information (explicitly or implicitly gathered) and domain ontology are used to handle heterogeneity and, consequently, provide users with more complete answers according to their current context of work.

4.4.1 Context in Data Integration

The problem of data integration is a challenge faced by applications that query across multiple autonomous and heterogeneous data sources. To help matters, contextual information may be employed to improve two important aspects in data integration: schema integration and query processing. However, dealing with contextual information entails a high development cost because several tasks (e.g. context acquisition and processing) must be addressed.

The use of context in data integration systems is quite different from other context-sensitive applications. Integrating heterogeneous data sources requires solving schematic and semantic conflicts which may arise at schema or instance-level [Kashyap & Sheth 1996, Goh 1997, Souza et al. 2006, Stefanidis et al. 2005]. Some of the metadata that describe the data sources may be used as contextual information (e.g. available query operator). Other contextual elements are perceived or inferred dynamically during the execution of a given process (e.g. in query processing, the availability of data sources is a rather important information that is obtained on the fly). Thus, in this section we discuss the use of context for two important areas in data integration: schema integration and query processing.

A general schema integration operation comprises a number of specific tasks that receive a set of different data source schemas, with varying structures and semantics, and produces an integrated schema with reconciled resulting elements. A schema integration process usually consists of the following tasks [Rahm & Bernstein 2001]: i) the preprocessing routine that translates the schemas into a common format and makes schema element names comparable; ii) the schema comparison which establishes the meaning of schema elements producing inter-schema mappings; and iii) the merging and restructuring tasks which group corresponding elements to generate the integrated schema. In schema integration processes, element names can have different meanings depending on the context in which they are related. Hence, CEs may improve the semantic interpretation of an entity by restricting or modifying the meaning of an element according to a specific context.

The other main issue in data integration systems is query processing. Such task usually includes the following steps: (i) query submission and analysis; (ii) relevant data sources' identification; (iii) query reformulation according to semantic mappings; (iv) query execution and results' integration; and (v) result presentation. Applying context reasoning to query processing

enriches the complete process as well as provides what has been called context-sensitive queries - those whose results depend on the context at the time of their submission [Stefanidis et al. 2005]. In this sense, when a user poses a query, all the surrounding CEs will be analyzed to avoid ambiguity, indicating the data that are really relevant to the user's specific situation. Besides, specific data conflicts arise mostly when query answers are assembled to produce a final result [Goh 1997]. Therefore, user profile, query model and interface, data sources' availability and semantic mappings are examples of CEs that may be used to contextualize queries, providing users with more meaningful results.

In order to better represent contextual information we have proposed a contextual ontology for data integration presented in the following.

4.4.2 A Contextual Ontology for Data Integration

CODI (**C**ontextual **O**ntology for **D**ata **I**ntegration) is an ontology for representing *context* according to the issues discussed previously. In order to establish the relevant contextual elements (CEs), at first we have identified which domain entities we needed to work with. A domain entity is anything in the real world that is relevant to describe the domain (e.g data sources, users and applications). We consider that CEs are used to characterize a given domain entity. Therefore, we determined six main domain entities around which we consider the CEs: *user*, *environment*, *data*, *procedure*, *association* and *application*. We present the domain entities' taxonomy in Figure 4.5. Considering such entities, we have identified which contextual elements are relevant to them, as shown in Figure 4.6. As a result, CODI is a conjunction of those domain entities and the CEs related to them.

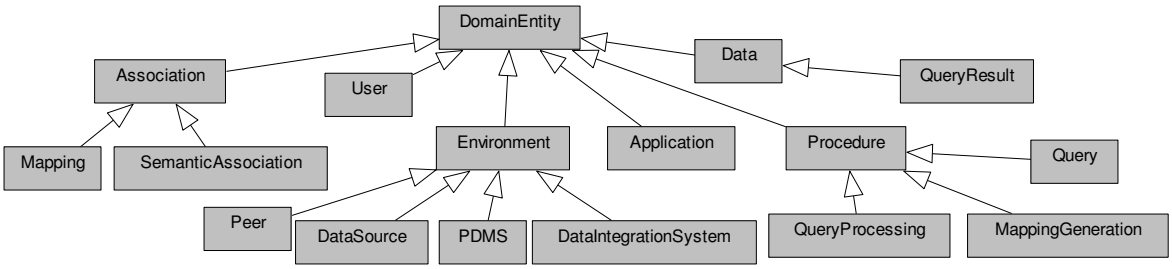


Figure 4.5 - The Domain Entities' taxonomy

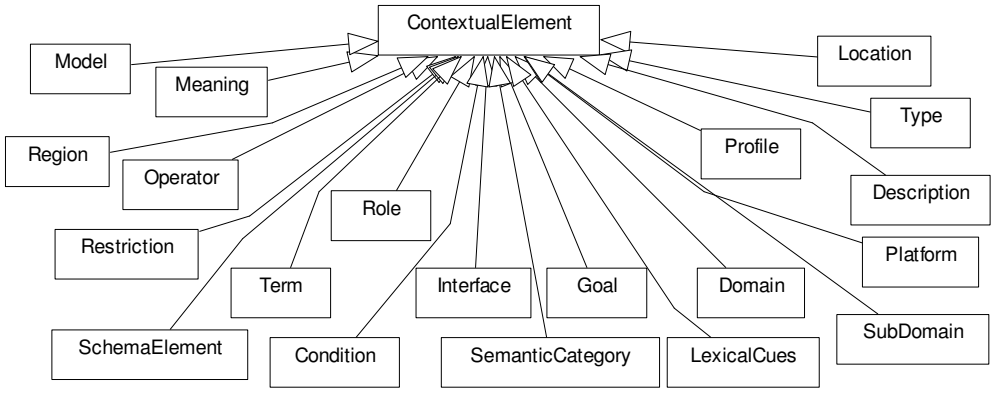


Figure 4.6 - An overview of CODI's Contextual Elements

CODI has been developed using Protégé 3.2[4]. For the sake of space, we have converted the diagrams to UML[5]. In addition, the contextual elements are shown in white and the domain entities in gray.

As an example, we present the *Data* domain entity, its relationship with other domain entities (*DataSource* and *SourceSchema*) and its contextual elements:

Data: data CEs are classified into *Schema Element Content* and *QueryResult*. A *Schema element content* is related to its *schema element* which is classified into *entity* and *attribute* CEs (Figure 4.7). *Schema element* constitutes one of the main concepts both to query processing and schema integration, since it is possible to infer semantic associations from its meaning (achieved when identifying its corresponding concept in the domain ontology). The *Query Result* represents results of individual queries as well as the final result obtained from the integration of several individual *Query Results*. Such result is presented to the user according to his/her CEs and intended level of detail.

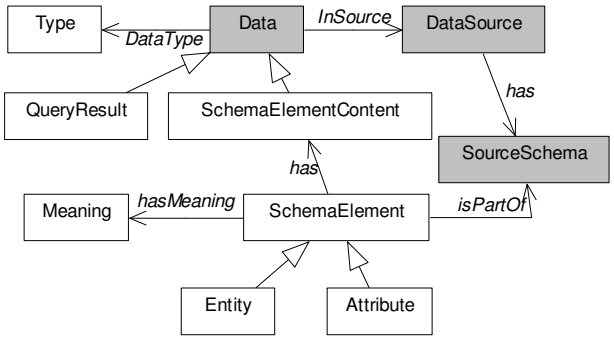


Figure 4.7 - CEs for the *Data* domain entity

Representing contextual elements using an ontology has various benefits. It provides concept subsumption, concept consistency and instance checking (including object properties checking). A contextual ontology also allows developers to define constraints and reasoning rules that may be used to derive other implicit context information. As an illustration, in query processing, we may use contextual elements to provide query expansion. Whenever elements 1 and 2 hold semantic relationships such as synonymy, meronymy or generalization to each other, an expansion rule may be created, as follows: "*Element1* $\rightarrow$ *Element2*" (i.e. element1 implies element2). In other words, if a query contains term *Element1*, then *Element2* is always considered as a candidate for expansion, depending on the kind of the semantic relationship (e.g. synonym or specialization).

4.4.3 Summary

CODI aims to assist the common tasks of a generic data integration process. This means that CODI represents CEs related to the entities involved within a data integration scenario from any knowledge domain. What differentiates CODI from other approaches is that the other ones lack important aspects that should be considered in data integration (e.g. procedure, environment and association) since they are usually restricted to specific integration processes and/or knowledge domains. CODI aims to structure entities and their CEs in such a way that they may be used for

[4] Protégé 3.2 beta-version, protege.stanford.edu/

[5] Unified Modeling Language

diverse processes, including schema integration [Belian 2008] and query reformulation [Souza 2009].

4.5 A Semantic Approach to Data Management in PDMS

Recently, Peer Data Management Systems (PDMS) came into the focus of research as a natural extension to distributed databases in the peer-to-peer (P2P) context [Herschel & Heese 2005, Tatarinov et al. 2003, Nejdl et al. 2002]. PDMS are P2P applications where each peer represents an autonomous data source and exports its entire data schema or only a portion of it. Such schemas, named exported schemas, represent the data to be shared with the other peers. Among those exported schemas, semantic mappings, i.e. correspondences between schema elements, are generated and maintained.

Mainly due to semantic heterogeneity, research on PDMS has considered the use of ontologies as a way of providing a domain reference and describing data sources in a uniform notation. In this sense, we propose an OPDMS, named SPEED, which adopts a semantic-based approach to assist relevant issues in peer data management [Pires 2009, Souza 2009]. The content shared by peers is represented through ontologies, which are used to group semantically related peers.

4.5.1 PDMS and Ontologies

PDMS realize their services over data from existing heterogeneous sources. As a result, they must be able to deal with different types of heterogeneity. The use of ontologies enables P2P interoperability at different levels of abstraction in terms of both peer matching (i.e. the process of comparing the content shared by two peers in order to determine the semantic similarity between them) and subsequent knowledge sharing. In this light, ontologies may be used as a way to enrich PDMS services and provide users with more complete results.

Xiao [Xiao 2006] has introduced the concept of OPDMS through two important issues: (i) ontologies are used in local sources as a uniform conceptual metadata representation; and (ii) ontology mappings are established between peers to allow query processing. We argue that ontologies may be used in a broader way to enhance PDMS services. Considering that, in this work, we propose an extension to the OPDMS description. Thus, we define an OPDMS as *a PDMS which is conceived for supporting dynamic ontology-based knowledge sharing, and this knowledge must be employed to improve its services*. Moreover, based on our analysis of the state-of-the-art on PDMS, we have identified a set of high-level requirements that an OPDMS should fulfill. Next, we briefly discuss each one:

(i) Exported schema representation: peers' metadata should be mapped onto an ontological description, using a common model;

(ii) Global conceptualization: a global ontology may be used to provide a high-level view over the heterogeneous peer schemas;

(iii) Support for mappings identification: an ontology may also be used to facilitate the identification of semantic mappings between peer exported schemas, i.e., between ontologies;

(iv) Support for query processing: query processing in a PDMS may use a global ontology in a two-fold way: a) as a high-level view of the sources; and b) as a terms' reference for query rewriting between peers. The former is concerned with query formulation, i.e., the user can formulate a query without specific knowledge of the different data sources stored in the peers. The latter is concerned with query rewriting, i.e., the query is rewritten into a target query over

other connected peers, according to a global ontology (its terms) and the defined semantic mappings;

(v) Semantic Index: since an ontology is employed as a global conceptualization of a group of peers, a semantic index can be built according to the main terms or categories of that ontology. Such index must enable efficient location of peers which are able to provide relevant data to a given query;

(vi) Semantic matchmaking capabilities: some peers implement a semantic matchmaker module for matching ontologies in order to find out which concepts match in different ontologies and (possibly) at which level. Such capability is mainly used for peer connectivity and the identification of semantic mappings.

A system should take into account the previous requirements not only to be considered an OPDMS, but also to take advantage of using ontologies for semantic enrichment. In order to fulfill those requirements, we propose an OPDMS architecture that will be described in the next section.

4.5.2 PDMS Architecture

In SPEED (**S**emantic **PEE**r-to-Peer **D**ata Management System) peers are organized according to their shared content. The system employs a mixed network topology (DHT [Stoica et al. 2001] and super-peer [Yang & Garcia-Molina 2003]) to exploit the strengths of both topologies. A DHT network is used to assist peers with common interests to find each other and form semantic communities. Within a community, peers are assembled in clusters, each one arranged in a super-peer topology.

As shown in Figure 4.8, three distinct types of peers are considered in the system: data peers, integration peers, and semantic peers. A **data peer** represents a data source sharing structured or semi-structured data with other data peers in the system. In Figure 4.8, I_1D_1 and I_1D_2 are examples of data peers. Data peers are grouped within **semantic clusters** according to their semantic interest. A **semantic interest** includes the peer **interest theme** and a **local peer ontology**. The interest theme is an abstract description of the peer semantic domain, whereas the local peer ontology describes the peer exported schema.

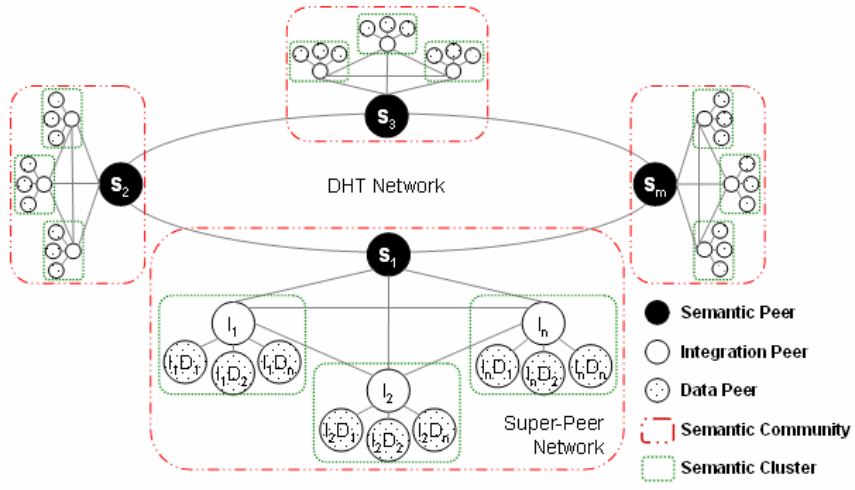


Figure 4.8 - Overview of SPEED architecture

Each semantic cluster has a special type of peer named **integration peer**. Actually, integration peers are data peers with higher availability, network bandwidth, processing power,

and storage capacity. Such peers are responsible for tasks like managing data peers' metadata, query processing, and data integration. In Figure 4.8, I_1 is the integration peer of the semantic cluster formed by the data peers I_1D_1, I_1D_2, and I_1D_n.

An integration peer maintains a **cluster ontology**, which is obtained through the merging of the local ontologies representing data peers' and integration peers' exported schemas. It acts as shared vocabulary inside a cluster, inter-relating semantically similar ontology concepts. Integration peers communicate with a **semantic peer**, which is responsible for storing and offering a **community ontology** containing elements of a particular knowledge domain. Semantic peers are responsible for managing integration peers' metadata. In Figure 1, S_1 is an example of a semantic peer. A set of clusters sharing related semantic interests forms a **semantic community**.

4.5.3 SPEED Community and Clusters

In SPEED, requesting peers are firstly assigned to a semantic community and then grouped within a semantic cluster. The semantic interest is of great importance since that the information contained on it (interest theme and local ontology) are used to associate requesting peers to adequate communities and clusters. A peer interest theme is an abstract description of its semantic domain.

Basically, the discovery of a semantic community is performed through the use of keywords. The search starts when the interest theme of a requesting peer is sent to an arbitrary semantic peer within the DHT network (Figure 4.9). Each semantic peer represents a distinct semantic domain (a semantic community). Semantic peers are searched according to a particular DHT protocol (e.g. Chord [Stoica et al. 2001]). In SPEED current version, a peer is able to participate in only one semantic community.

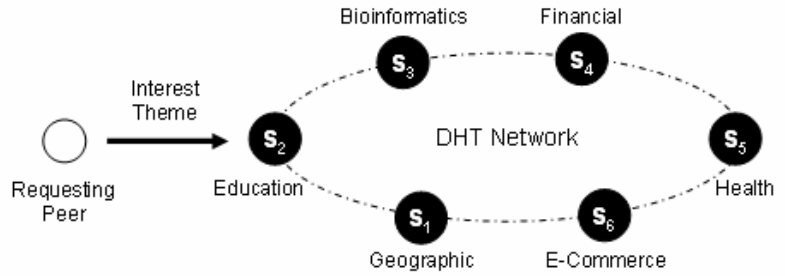


Figure 4.9 - Community discovery through keyword search

Once the semantic community has been discovered, the requesting peer must find out an appropriate semantic cluster. In this sense, differently from the community discovery, cluster discovery is performed through ontology matching (see Figure 4.10). The semantic matchmaker module performs a matching between the local ontology (requesting peer) and the cluster ontology (integration peer), producing a similarity degree between them.

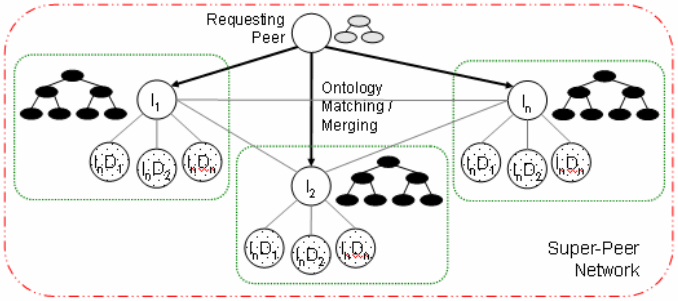


Figure 4.10 - Cluster discovery (ontology matching) and cluster formation (ontology merging)**.**

The requesting peer is relevant for the cluster if the semantic ontology matchmaker produces a value higher than a *cluster threshold*. If the comparison result is higher than the cluster threshold, then the requesting peer will participate in that cluster. Otherwise, it is addressed to another integration peer within the same semantic community. If the comparison result is lower than the cluster threshold for all existing clusters, then the requesting peer starts the formation of a new cluster. In general, a requesting peer is initially connected as a data peer. However, in the case where it is the first peer in a cluster, it is connected as an integration peer. Thus, the cluster ontology is created when the first data peer is connected to the cluster. As long as other data peers join the cluster, the integration peer can extend the cluster ontology by adding new concepts and properties through ontology merging.

4.5.4 Summary

We have presented SPEED, an OPDMS that makes use of ontologies to improve the quality of its services, such as peer connectivity. To this end, ontologies are employed in various aspects, namely: uniform representation of peers' contents, global conceptualization, support for query processing, creation and maintenance of semantic indexes, support for mappings identification, and semantic matchmaking. A distinguishing characteristic of SPEED is that peers are organized in a two-tier architecture. A broader grouping level (communities) is used to allow peers to share content associated to related semantic domains. Such level eases resource discovery by assisting peers to efficiently find other related ones. Additionally, a finer grouping level (clusters) is employed to improve the generation of semantic mappings and enhance query results. Finally, SPEED architecture and services have been designed according to the OPDMS requirements in order to completely fulfill all of them. Two PhD theses are being developed within this project [Pires 2009, Souza 2009].

4.6 Main Results

The main results related to semantic issues can be summarised as:

- a context-oriented model and a generic context manager
- a contextual ontology for data integration
- a semantic-based PDMS architecture

As an ongoing project other topics are being developed related to the formation of clusters with semantic related peers and a query reformulation process.

In what follows we present the PhD thesis associated to this research area and the related publications.

4.6.1 PhD and MSc Theses

In this research topic, two PhD and one MSc theses were already concluded and there are three ongoing PhD theses. They are all listed hereafter.

[Siebra 2007] Sandra de Albuquerque Siebra. Analysing Participants' Interactions in Collaborative Learning Environments, PhD, 2007

[Petry 2007] Helô Petry. ICARE: Um Sistema de Recomendação de Especialistas Sensível a Contexto, MSc, 2007

[Belian 2008] Rosalie Barreto Belian. A Context-based Name Resolution Process for Schema Integration, PhD, 2008

[Vieira 2008] Vaninha Vieira dos Santos. A Domain Independent Approach for Managing Contextual Elements, PhD, 2004-2008 (ongoing)

[Souza 2009] Damires Yluska de Souza Fernandes. Semantic-based Query Reformulation for PDMS, PhD, 2005-2009 (ongoing)

[Pires 2009] Carlos Eduardo Santos Pires. Semantic-based Approach for Peer Clustering in a Peer Data Management System, PhD, 2005-2009 (ongoing)

4.6.2 Publications

The results already obtained in this research area were published in two journal papers, *International Journal of Intelligent Information Technologies* and *Journal of the Brazilian Computer Society*, and another is in evaluation process for the *Revue d'Intelligence Artificielle*. In addition, there are two publications in the main conference on the context community, *International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT)*, and some others in specific workshops.

[Vieira et al. 2008] Vieira, V., Brezillon, P., Salgado, A. C., Tedesco, P. C. A. R. A Context-Oriented Model for Domain-Independent Context Management, *Revue d'Intelligence Artificielle*, edited by Lavoisier, submitted to evaluation, 2008

[Siebra et al. 2007] Siebra, S. A., Salgado, A. C., Tedesco, P. C. A. R. A Contextualized Learning Interaction Memory, *Journal of the Brazilian Computer Society*, Vol.13, N.3, 2007, p.51 – 66

[Vieira et al. 2007a] Vieira, V., Tedesco, P. C. A. R., Salgado, A. C., Brezillon, P. Investigating the Specifics of Contextual Elements Management: The CEManTIKA Approach, In Proc. of the *6th International and Interdisciplinary Conference on Modeling and Using Context*, Lecture Notes in Artificial Intelligence – LNAI, v.4635, Roskilde, Denmark, 2007. p.493 - 506

[Vieira et al. 2007b] Vieira, V., Brezillon, P., Salgado, A. C., Tedesco, P. C. A. R. Towards a Generic Contextual Elements Model to Support Context Management, In Proc. of the *4th International Workshop on Modeling and Reasoning in Context (MRC 2007)*, Computer Science Research Report Vol.112, Roskilde, Denmark, 2007, p.49 - 60

[Cruz et al. 2007] Cruz, E., Vieira, V., Almeida, E.S., Meira, S.L., Salgado, A.C., Brezillon, P. Modeling Context in Software Reuse, In Proc. of the *4th International Workshop on Modeling and Reasoning in Context (MRC 2007)*, Computer Science Research Report Vol.112, Roskilde, Denmark, 2007, p.89 – 102

[Cardoso et al. 2006] Cardoso, R. C., Souza, F. F., Salgado, A. C. Retrieving Specific Domain Information from the Web through Ontologies, *International Journal of Intelligent Information Technologies*, edited by Idea Group Inc., Vol.2, N.3, 2006, p.56 - 71

[Souza et al. 2006] Souza, D., Salgado, A. C., Tedesco, P. C. A. R. Towards a Context Ontology for Geospatial Data Integration, In Proc. of the *2nd International Workshop on Semantic-based Geographical Information Systems (SeBGIS'06)*, Lecture Notes in Computer Science – LNCS, v. 4278, Montpellier, France, 2006, p.1576 – 1585

[Petry et al. 2006] Petry, H., Vieira, V., Tedesco, P.C.A.R., Salgado, A.C. Um Sistema de Recomendação de Especialistas Sensível ao Contexto para Apoio à Colaboração Informal, In Proc. of the *Simpósio Brasileiro de Sistemas Colaborativos*, Natal, 2006. p.38 – 47

[Siebra et al. 2006] Siebra, S. A., Tedesco, P.C.A.R., Salgado, A.C. A Process for User Interaction Analysis in Collaborative Environments In Proc. of the *Simpósio Brasileiro de Sistemas Colaborativos*, Natal, 2006. p.19 - 28

[Siebra et al. 2005a] Siebra, S. A., Salgado, A. C., Tedesco, P. C. A. R., Brezillon, P. A Learning Interaction Memory using Contextual Information, In Proc. of the *Context and Work Group*

Workshop, CEUR Workshop Proceedings (http://CEUR-WS.org/Vol-133/), Paris, France, 2005, p.1 - 12

[Siebra et al. 2005b] Siebra, S. A., Salgado, A. C., Tedesco, P. C. A. R., Brezillon, P. Identifying the Interaction Context in CSCLE, In Proc. of the *5th International and Interdisciplinary Conference on Modeling and Using Context (Context05)*, Lecture Notes in Artificial Intelligence – LNAI, v.3554, Paris, France, 2005, p.464 – 475

[Vieira et al. 2005a] Vieira, V., Salgado, A. C., Tedesco, P. C. A. R. Towards an Ontology for Context Representation in Groupware, In Proc. of the *11th International Workshop on Groupware (CRIWG 2005)*, Lecture Notes in Computer Science – LNCS, v.3706, Porto de Galinhas, Brazil, 2005, p.367 – 375

[Vieira et al. 2005b] Vieira, V., Tedesco, P.C.A.R., Salgado, A.C. Representação de Contextos em Ambientes Colaborativos Usando Ontologia, In Proc. of the *Wokshop Brasileiro de Tecnologias para Colaboração (WCSCW)*, Juiz De Fora, 2005, p.721 - 730

[Cardoso et al. 2005] Cardoso, R. C., Souza, F. F., Salgado, A. C. Using Ontologies to Prospect Offers on the Web, In Proc. of the *7th International Conference on Enterprise Information Systems (ICEIS)*, Miami, USA, 2005, p.200 - 207

4.7 References

[Bézivin 1998] Bézivin, J. Who´s Afraid of Ontologies. In Proc. of the *Model Engineering, Methods and Tools Integration with CDIF Workshop (OOPSLA'98)*, Vancouver, Canada, 1998

[Brézillon 2003] Brézillon P. Context Dynamic and Explanation in Contextual Graphs. In Proc. of the *4th International and Interdisciplinary Conference (CONTEXT 2003)*, Stanford, USA, 2003, p. 94-106

[Brézillon & Pomerol 1999] Brézillon, P., Pomerol, J.-C. Contextual Knowledge Sharing and Cooperation in Intelligent Assistant Systems, *Le Travail Humain*, Paris, Vol. 62, N. 3, 1999, p. 223-246.

[Chandrasekaran et al. 1999] Chandrasekaran, B., Josephson, J.R., Benjamins, V.R. What Are Ontologies, and Why Do We Need Them? *IEEE Intelligent Systems,* Vol.14, N.1, 1999, p. 20-26

[Dey & Abowd 2000] Dey, A. K., Abowd, G. D. Towards a Better Understanding of Context and Context-awareness, In Proc. of the *Workshop on the What, Who, Where, When, Why and How of Context-awareness (CHI 2000)*, The Hague, The Netherlands, 2000, p. 1-6

[Goh 1997] Goh, C. Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Systems, *PhD. Thesis*, MIT, 1997

[Herschel & Heese 2005] Herschel, S., Heese, R. 2005. Humboldt Discoverer: A Semantic P2P index for PDMS, In Proc. of the *International Workshop Data Integration and the Semantic Web (DisWeb)*, Porto, Portugal, 2005

[Kashyap & Sheth 1996] Kashyap, V., Sheth, A. Semantic and Schematic Similarities between Database Objects: a Context-based Approach, *The VLDB Journal*, Vol.5, N.4, 1996, p. 276-304

[Nejdl et al. 2002] Nejdl, W., Wolf, B., Qu, C., Decker, S., Sintek, M., Naeve, A., Nilsson, M., Palmér, M., and Risch, T. Edutella: a P2P Networking Infrastructure Based on RDF, In Proc. of the *11th Int. World Wide Conference*, Hawaii, USA, 2002, p. 604-615.

[Noy & McGuiness 2001] Noy, N., McGuiness, D.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05, March 2001. (also available on www.ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noymcguinness.html, last access February 2008)

[Rahm & Bernstein 2001] Rahm, E., Bernstein, P.. A Survey of Approaches to Automatic Schema Matching, *The VLDB Journal*, Vol.10, N.4, 2001, p. 334-350

[Stefanidis et al. 2005] Stefanidis K., Pitoura E., Vassiliadis P. On Supporting Context-Aware Preferences in Relational Database Systems, In Proc. of the *1st International Workshop on Managing Context Information in Mobile and Pervasive Environments (MCMP'2005)*, Ayia Napa, Cyprus, 2005

[Stoica et al. 2001] Stoica, I., Morris, R., Karger, D., Kaashoek, M. F., Balakrishnan, H. Chord: a Scalable Peer-to-Peer Lookup Service for Internet Applications. In Proc. of the *ACM SIGCOMM 2001*, San Diego, USA, 2001, p. 149-160.

[Tatarinov et al. 2003] Tatarinov, I., Ives, Z., Madhavan, J., Halevy, A., Suciu, D., Dalvi, N., Dong, X., Kadiyska, Y. Miklau, G., Mork, P. The Piazza Peer Data Management Project, In Proc. of the *ACM SIGMOD Record*, Vol.32, N. 3, 2003, p. 47-52

[Wache et al. 2001] Wache H., Voegele T., Visser U., Stuckenschmidt H. Ontology-based Integration of Information – A Survey of Existing Approaches, In Proc. of the *IJCAI-01 Workshop: Ontologies and Information Sharing*, Seattle, USA, 2001, p. 108-117

[Xiao 2006] Xiao, H. Query Processing for Heterogeneous Data Integration using Ontologies, *Ph.D. Thesis*, University of Illinois, Chicago, USA, 2006

[Yang & Garcia-Molina 2003] Yang, B., Garcia-Molina, H. Designing a Super-Peer Network. In Proc. of the *19th International Conference on Data Engineering*, Bangalore, India, 2003, p. 49-60

[Ziegler & Dittrich 2004] Ziegler P., Dittrich, K. : Three decades of data integration: all problems solved? In Proc. of the *18th IFIP World Computing Congress*, Toulouse, France, 2004, p.3-12

CHAPTER 5

Other Activities

5.1 Research Activities

In addition to the three areas previously presented there are other important research areas that I have been work and it is also important to speak about.

The first one was Multimedia Databases which continued the work initiated during my PhD studies. We have proposed a data type manager allowing the definition of new data types along with its corresponding operators [Sa 1991], an extension of the relational model to deal with multimedia features [Cruz 1992] and some features to be included in a multimedia station [Santos 1995], corresponding respectively to three MSc theses. The results were published in national conferences [Sá & Salgado 1991, Santos et al. 1996b, Santos et al. 1996b] and one paper in a Brazilian journal [Cruz & Salgado 1991].

Another area we have been working on is Cooperative Systems. In this area, our first interests were database cooperative transactions and object versioning, each subject associated to two MSc theses [Santos 1992a, Santos 1992b] and to some publications [Santos & Salgado 1992a, Santos & Salgado 1992b, Santos et al. 1993]. Another contribution was the definition of a communication manager for cooperative environment along with its graphical interface resulting in two other MSc theses [Aguiar 1992, Mesquita 1996] and to the following publications [Aguiar & Salgado 1992, Aguiar & Salgado 1993, Aguiar & Salgado 1994, Mesquita & Salgado 1996]. Afterwards, I participated in an iberoamerican multi-institutional project financed by CYTED, a Spanish institution. The objective of this research project was to generate a tool for meeting preparation (pre-meeting) based on *SISCO*, our proposed argumentation model [Bellasai et al. 1995, Borges et al. 1999]. More recently, along with a PhD student, we have proposed the analysis of participants' interactions in Collaborative Learning environments using contextual information and a data warehouse as learning memory [Siebra 2007, Siebra et al. 2005a, Siebra et al. 2005b, Siebra et al. 2006, Siebra et al. 2007].

We have also participated, along with other Information Retrieval researchers, in the specification and implementation of *Radix*[6] [Gonçalves et al. 1997a, Gonçalves et al. 1997b, Cardoza et al. 1998], a search engine which indexed all Brazilian pages (.br) and was largely used in Brazil at this time. Some master theses were developed in the context of this project: a pre-processing analysis with query expansion [Guerra 2001], a block-based strategy to inverted indexes [Miranda 2003], an strategy to search engines' database updates based on classifiers [Barbosa 2003, Barbosa et al. 2002, Barbosa et al. 2003] and a mining process in search engines

[6] www.radix.com.br

databases [Arakaki 2003]. This project had financial support of a Brazilian investment bank. The start- up was created in 1999 and was sold in 2001. Radix is currently the search tool of the IG (Internet Group)7 portal in Brazil.

It is worth noting that in all these research projects several undergraduate students have participated in what we call in Brazil the 'Scientific Initiation' program. It is a national scholarship program that motivates undergraduate students to participate in research project. In this light, I have supervised at least two of these students in each project. The good issue is that some of them have continued in the MSc and PhD studies.

5.2 Additional Activities

In parallel with the research activities, I have actively participated in the consolidation and growth of the Informatics Department at UFPE and the consequent creation of the Center for Informatics (a highest level in the University hierarchy). During this time, I had several administrative positions, i.e. coordinator of Computer Science undergraduate studies (1992-1996), head of Informatics Department (1997 and 1999-2001), and finally director of the Center for Informatics (2001-2005).

In 1996 I was among the founders, along with seven other colleagues, of C.E.S.A.R[8] (Center for Studies and Advanced Systems of Recife), a non-profitable private institution. C.E.S.A.R has the mission of creating products, processes, services and innovative companies in Information Technology. I am currently the chairwoman of its administrative board.

5.3 Main Results

The two main results obtained in the additional research area presented in Section 5.1 were: *Radix*, a real product and largely used in the internet, and *SISCO* a tool for meeting preparation.

In this section we will also present the supervised MSc theses and the some publications.

5.3.1 MSc Theses

Ten concluded MSc theses are presented in chronological order.

[Sá 1991] Jessica Barros De Sá. Get: Um Gerenciador de Tipos de Dados, MSc, 1991

[Santos 1992a] Afra Maria Barbosa Martiniano Dos Santos. Gerenciamento de Versões de Objetos em um Ambiente de Banco de Dados: Análise e Proposta, MSc,1992

[Santos 1992b] Maurilucio Martiniano dos Santos. Um Modelo para o Tratamento de Transações Longas em Ambiente de Bancos de Dados, MSc, 1992

[Cruz 1992] Maria Lencastre P. Menezes E Cruz. Br+ : Um Modelo Dinâmico de Dados para um Ambiente Multimidia, MSc, 1992

[Aguiar 1993] Carlos Augusto Teixeira De Aguiar. Um Ambiente de Suporte ao Trabalho Cooperativo, MSc, 1993

[7] busca.igbusca.com.br

[8] www.cesar.org.br

[Santos 1995] Marizete Silva Santos. Estações Multimídia: Projeto e Inclusão de Novos Recursos, MSc, 1995

[Mesquita 1996] Cláudia do Socorro Ferreira Mesquita. Icaro - Uma Interface de Comunicação Para Ambientes Cooperativos, MSc, 1996

[Guerra 2001] Marcela Fontes Lima Guerra. SIPEC - Um Sistema Interativo de Pré-processamento e Expansão de Consulta para Recuperação de Informações na Web, MSc, 2001

[Miranda 2003] Oscar Gomes de Miranda. VIF - Uma Estrutura de Índice Invertido em Blocos Baseada em uma B+-tree, MSc, 2003

[Barbosa 2003] Luciano de Andrade Barbosa. Uma Proposta para a Atualização de Bases de Dados em Engenhos de Busca Utilizando Classificadores, MSc, 2003

[Arakaki 2003] Eduardo Massao Arakaki. Sistema de Análise de Dados de Acesso a um Engenho de Busca, MSc, 2003

5.3.2 Publications

Among these publications I would like to highlight one journal paper in *Decision Support Systems*, three Brazilian journal papers and the following international conferences: *INET*, *CAISE*, *CONTEXT* and *IRMA*.

[Siebra et al. 2007] Siebra, S.A., Salgado, A.C., Tedesco, P.C.A.R. A Contextualized Learning Interaction Memory, *Journal of the Brazilian Computer Society*, Vol.13, N.3, 2007, p.51 – 66

[Siebra et al. 2006] Siebra, S. A., Tedesco, P.C.A.R., Salgado, A.C. A Process for User Interaction Analysis in Collaborative Environments In Proc. of the *Simpósio Brasileiro de Sistemas Colaborativos*, Natal, 2006. p.19 - 28

[Siebra et al. 2005b] Siebra, S. A., Salgado, A. C., Tedesco, P. C. A. R., Brezillon, P. Identifying the Interaction Context in CSCLE, In Proc. of the *5th International and Interdisciplinary Conference on Modeling and Using Context (Context05)*, Lecture Notes in Artificial Intelligence – LNAI, v.3554, Paris, France, 2005, p.464 – 475

[Siebra et al. 2005a] Siebra, S. A., Salgado, A. C., Tedesco, P. C. A. R., Brezillon, P. A Learning Interaction Memory using Contextual Information, In Proc. of the *Context and Work Group Workshop*, CEUR Workshop Proceedings (http://CEUR-WS.org/Vol-133/), Paris, France, 2005, p.1 - 12

[Barbosa et al. 2003] Barbosa, L. A., Ramalho, F. S., T. Neto, M. C., Salgado, A. C. Dynamic Indexing of Information in the Web: the Case of News Sites, In Proc. of the *14th International Resources Management Association International Conference(IRMA)*, Idea Group Publishing, Philadelphia, USA, 2003. p.279 – 282

[Barbosa et al. 2002] Barbosa, L. A., Ramalho, F. S., Salgado, A. C. Indexando Informação Dinâmica na Web: o Caso dos Sites de Notícias (Indexing Dynamic Information on the Web: the case of news sites). *Revista de Informática Teórica e Aplicada*, edited by Federal University of Rio Grande do Sul, Vol.9, N.3, 2002, p.59 - 72

[Borges et al. 1999] Borges, M. R. S., Pino, J., Salgado, A. C., Fuller, D. Key Issues in the Design of an Asynchronous System to Support Meeting Preparation, *Decision Support Systems*, edited by Elsevier, v.27, n.3, p.271 - 289, 1999

[Cardoza et al. 1998] Cardoza, J., Goncalves, P. F., Lima, A., Pereira, L., Tercero, C., Meira, S. L., Salgado, A. C., Silva, F. Q. B. A Framework for Developing Information Indexing, Filtering and Searching Tools in the Internet, In Proc. of the *8th Annual Conference of the Internet Society (INET)*, Geneva, Switzerland, 1998 (also available at http://www.isoc.org/inet98/proceedings/1c/1c_3.htm, last access March 2008)

[Gonçalves et al. 1997a] Gonçalves, P. F., Salgado, A. C., Meira, S. L. A Distributed Mobile-Code Architecture for Information Indexing and Searching, In Proc. of the *7th Annual Conference of the Internet Society (INET)*, Kuala Lumpur, Malaysia, 1997 (also available at http://www.isoc.org/inet97/proceedings/A7/A7_2.htm , last access March 2008)

[Gonçalves et al. 1997b] Gonçalves, P. F., Salgado, A. C., Meira, S. L. Digital Neighborhoods: Partitioning the Web for Information Indexing and Searching, In Proc. of the *9th Conference on Advanced Information Systems Engineering (CAISE)*, Barcelona, Spain, 1997, p.289 - 302

[Santos et al. 1996b] Santos, M., Souza, F.F., Salgado, A.C. Princípios de Design na Construção de Interfaces para Sistemas Multimídia, In Proc. of the *2nd Workshop em Sistemas Hipermídia*, Fortaleza, Brazil, 1996

[Santos et al. 1996a] Santos, M., Souza, F.F., Salgado, A.C. Estações Multimídia de Atendimento: Uma Abordagem Ergonômica, In Proc. of the *22nd Confêrencia Latino-Americana de Informática (PANEL)*, Bogotá, Colombia, 1996

[Bellasai et al. 1995] Bellasai, G., Borges, M.R.S., Fuller, D., Pino, J., Salgado, A.C. Sisco - A Tool to Improve Meeting Productivity, In Proc. of the *1st CYTED-RITOS International Workshop on Groupware*, Lisbon, Portugal, 1995, p.149 – 161

[Mesquita & Salgado 1995] Mesquita, C., Salgado, A.C. ITG-A Multi-user Graphical Communication Tool, In Proc. of the *15th International Conference of the Chilean Computer Science Society*, Arica, Chile, 1995, p.284 - 293

[Aguiar & Salgado 1994] Aguiar, C.A.T., Salgado, A.C. The Management of a Cooperative Environment, *Computer Science 2 - Research and Applications*, edited by PLENUM PRESS, 1994, p. 469-478.

[Santos et al. 1993] Santos, A.B.M., Santos, M.M., Salgado, A.C. Uma Abordagem para o Tratamento de Transações Não Convencionais Utilizando Versões de Objetos, In Proc. of the *14th Conferência Latino-Americana de Informática*, Buenos Aires, Argentina, 1993. p.125 - 139

[Aguiar & Salgado 1993] Aguiar, C.A.T., Salgado, A.C. Um Gerenciador de Comunicação Para Ambientes Cooperativos, In Proc. of the *13th Congresso da Sociedade Brasileira de Computação*, Florianópolis, Brazil, 1993, p.431 – 445

[Aguiar & Salgado 1992] Aguiar, C.A.T., Salgado, A.C. Um Ambiente de Suporte ao Trabalho Cooperativo, In Proc. of the *18th Conferência Latino-Americana de Informática*, Las Palmas de Gran Canarias, Espanha, 1992, p.17 – 25

[Santos & Salgado 1992a] Santos, A.B.M., Salgado, A.C. Object Version Management in a Database Environment, In Proc. of the *12th International Conference of the Chilean Computer Science Society*, Santiago/Chile, 1992, p.27 – 38

[Santos & Salgado 1992b] Santos, M.M., Salgado, A.C. Um Mecanismo para o Gerenciamento de Transações Compostas, In Proc. of the *12th Congresso da Sociedade Brasileira de Computação*, Rio de Janeiro, Brazil, 1992, p.137 - 150

[Sá & Salgado 1991] Sa, J. B., Salgado, A. C. Get: Um Gerenciador Dinâmico de Tipos de Dados, In Proc. of the *6th Simpósio Brasileiro de Banco de Dados*, Manaus, Brazil, 1991, p.223 - 237

[Salgado & Cruz 1991] Salgado, A.C., Cruz, M.L. P. M. Br+: Um Modelo Dinâmico de Dados (Br+: A Dynamic Data Model). Revista de Informática Teórica e Aplicada, edited by Federal University of Rio Grande do Sul, Vol.1, N.4, 1991, p.23 – 46

CHAPTER 6

Conclusion and Perspectives

6.1 Summary

This report presents the main areas I have been working since my PhD studies, concluded in 1988 in Sophia-Antipolis, France. In fact, it contains our contributions in three main areas, i.e., Geographical Databases, Data Integration and Semantic Issues in PDMS. I also briefly presented other research areas I participated and administrative activities I assumed. I tried to show an overview of the main contributions and results obtained in each theme, along with the publications and the theses supervised.

Among these areas, I am currently interested in semantic issues in both data integration and PDMS. My current project proposes a semantic approach to peer clustering in PDMS to facilitate query routing in this environment. There are still a lot of challenges and future trends that are discussed in the following section.

6.2 Future Trends

In this section I will present some open issues in data integration and peer data management I am currently interested. I would finish this perspective analysis saying that considering web information systems there are two keywords of my research interest which are still not completely explored and need a lot of attention in the future: semantics and quality. In what follows I will present some open issues in data integration environments and discuss the challenges of our semantic-based approach to PDMS.

6.2.1 Open Issues in Data Integration Environments

There is a growing need for data integration in several environments: Web Information Systems (WIS), Peer Data Management Systems (PDMS), Enterprise Information Integration (EII), Personal Information Management (PIM), to say some. In this sense, even if a lot of work has already been done it is important to highlight some future (or ongoing) trends in data integration systems (in all mentioned environments):

- data sources' description: are metadata enough?
- entity resolution: what data objects refer to the same real-world entity?
- 'automatic' generation and evolution of schema mappings: where and when must the user intervene?
- reformulation and query processing: what are the query plans?

- quality criteria: how to measure data and information quality?
- scalability and performance: what about the dynamic environments (PDMS)?

Mapping evolution, mainly considering dynamic environments, and information quality criteria, mainly quality of the integrated schema, are the issues I am particularly interested and in which I have been working lately.

6.2.2 Semantic Peer Data Management

Peer data management systems are a natural extension of data integration systems, and thus have the same open issues previously discussed. Particularly, our proposal is a semantic approach to peer data management in which the idea is to group semantically related peers in clusters and communities (set of clusters of the same domain). The peer schemas are represented by ontologies that are aligned according to the corresponding domain ontology. In addition to the data integration open issues previously listed, we can identify other specific problems we have to deal with:

Identification of semantically similar peers to form clusters dynamically
The idea is that clusters are formed when peers connects to the system. As schema peers are represented by ontologies, we have to match the arriving peer with the cluster ontology to identify the cluster of the new peer. The cluster ontology reflects the content of all peer schemas it contains. We are analyzing the existing ontology matchers to identify the one that responds better to our needs. These matchers do not offer an overall similarity measure and we are working on the proposition of such metric considering the semantics of each term in the peer schema.

Identification of semantic neighbour clusters
Clusters that are semantically related must be identified as 'neighbours'. This is very important for the definition of schema mappings and for query routing strategies. The idea is to redirect queries to the semantically closest peers. Peer neighbours are the ones that have the semantic similarity measure above a pre-defined threshold.

Contextual-based query rewriting between neighbour clusters
As peer schemas are represented by ontologies and we are using OWL as representation language, we are currently working on the definition of a schema mapping representation language based on Description Logics. These mappings will be used in the process of query rewriting considering also contextual information. Information about the users, the peers related to the query answering, their respective clusters and corresponding neighbours, and the current state of the network, are examples of contextual information that may be considered in the query rewriting process. The idea is to improve the query processing and the quality of its result.

Mapping evolution
As PDMS is a very dynamic environment and schemas are represented by ontologies, the evolution problem includes the evolution of the involved ontologies (peers and clusters). The ontology evolution problem includes not only the maintenance of the ontology consistency but also of the corresponding schema mappings. Particularly in our SPEED architecture, the cluster ontology will evolve frequently because of the dynamicity of peer connectivity and this will provoke the frequent evolution of mappings between neighbour clusters. This requires robust change management algorithms.

6.2.3 Semantics and Quality

Semantics

The need of Semantic Web features is obvious but most of the proposed ideas remain largely unrealized. We are still far from retrieving the right data we would like using more complex queries than keyword-based ones. In this sense, I would also highlight two key concepts: ontology and context.

It is clear that we need more semantics to handle information mainly in heterogeneous environments where we have to share data and thus to reconcile terminologies. Actually, web sources do not have enough metadata to allow this task and the use of *ontologies* as a common conceptualization is essential. However, existing ontology matchers do not consider in their algorithms all the necessary semantic information to measure similarity.

In the last years, we saw the growth of context-based or context-sensitive systems. The concept of *context* is larger than space and time information usually employed in ubiquitous systems. In fact, contextual information about users, applications, environments, data and relationships, may also be employed to improve data management processes. The open issues are how to automatically acquire and manage contextual information.

To deal with semantic issues a huge interaction with Artificial Intelligence (AI) discipline is crucial. We need to use techniques of knowledge representation (including ontologies and contextual information), logics (causal, temporal and probabilistic), rules and inference, description logics, machine learning, only to refer some of them.

Quality

All the challenging data integration problems related to data uncertainty, inconsistency, incompleteness, provenance and trust are resumed in one word: quality. Once identified the relevant data quality criteria, it is necessary to identify metrics to evaluate them. In this sense, some works have been proposed. However, as a multi-dimensional concept it is not so simple to define usable data quality metrics.

Finally, we see two applications domains as the ones that need more semantics and more quality in their data managements systems: e-science, mainly life and environment sciences, and education, particularly e-learning.

6.3 Conclusion

The sabbatical year in France, and the preparation of this HDR report, can be seen as a review of all my research activities and also as a starting point to a new research phase. My current projects are being developed in cooperation with international partners and with the participation of some of my previous students. I hope that the future projects will go in the same direction in a collaborative and open way. This will allow to more easily facing up to the incoming challenges.

ANNEXE 1

Collaborative Projects

1. Project Name AMSUD

Title Evolution and quality management in dynamic data integration systems

Financial support AMSUD/STIC (CNRS, Capes, DICYT)

Period 2007-2009

Partners Facultad de Ingeniería – Universidad de la República - coordinator
Universidade Federal de Pernambuco
Universidade Federal do Ceará
Université de Versailles Saint-Quentin en Yvelynes
Laboratoire LSIS, Université Paul Cézanne (Aix-Marseille)

Role Member

Abstract The project overall objective is the development of techniques, algorithms and tools to provide support for the evolution and quality management in data integration systems. Different types of data integration systems will be considered for this project, including well structured ones, as mediation systems, and less structured ones as peer data management systems. Besides the technical and scientific results, this project will be of fundamental importance for strengthening the collaboration among the partners and for fostering new partnerships.

Main results **-** specification and implementation of schema quality criteria

Number of involved students

PhD	2
MSc	1
Undergraduate	2

Publications

Journals: 2 submitted for evaluation
International Conferences and workshops: 3

2. Project Name SPEED

Title Semantic Peer-to-Peer Data Management System

Financial support CNPq

Period 2008-2010

Partners Universidade Federal de Pernambuco
Universidade Federal do Ceará
Université de Versailles Saint-Quentin en Yvelynes

Role Cooordinator

Abstract Data management in P2P systems is a challenging and difficult problem considering the excessive number of peers, their autonomous nature and the potential heterogeneity of their schemas. A PDMS is considered as an evolution of the traditional data integration systems, in such a way that the notion of a single mediation schema is replaced by a set of semantic mappings between peers' schemas. We have proposed a semantic-based PDMS architecture where peers are clustered according to their semantic interest. Ontologies are used at different levels as a semantic common model. Our interest is centred on the definition of a clustering strategy of data peers, on the definition of an ontology-based mapping representation and on the corresponding query processing.

Main results
- an ontology based PDMS architecture
- a meta-ontology for data integration

Number of involved students

PhD	4 (3 ongoing)
MSc	1
Undergraduate	4

Publications

International Conferences and workshops: 1
National Conferences: 1

3. Project Name INTEGRA

Title Information Integration in Heterogeneous Environment: architectures, models and implementations

Financial support CNPq

Period 2003-2005

Partners Universidade Federal do Ceará
Université de Versailles Saint-Quentin en Yvelynes

Role Cooordinator

Abstract The problem of integrating data from heterogeneous and distributed data sources in a mediator-based environment consists in defining mappings between the mediation schema and the data source schemas, and providing a uniform view of the involved data sources. Our interest is focused on the maintenance of the mediation schema to reflect the evolution of the data source schemas, on the quality of the generated mediation schema and on the query reformulation process. A mediator-based data integration system prototype was implemented to validate the proposed approaches.

Main results
- a conceptual model to represent XML Schema entities: X-Entity Model
- a query reformulation process
- specification and implementation of schema quality criteria
- a data integration system prototype

Number of involved students

PhD	1
MSc	5
Undergraduate	4

Publications

Journals: 2
International Conferences and workshops: 8
National Conferences: 3

4. Project Name SISCO

Title Computer Supported Cooperative Systems: an Approach to Meeting Preparation

Financial support CYTED/RITOS (Ibero-american Support)

Period 1995-1998

Partners Universidade Federal do Rio de Janeiro, Brasil
Universidade Federal de Pernambuco, Brasil
Universidad de Chile, Chile
Pontifícia Universidad de Católica de Chile, Chile
Universidad Católica de Assuntión, Paraguay
Instituto de Engenharia de Sistemas e Computadores (INESC), Portugal
Centro de Investigación Científica y de Educación Superior de Ensenada, Mexico

Role Member

Abstract Cooperative Systems is an area that examines how computers can assist several people working together. We were particularly interested in creating a cooperative environment to improve meetings productivity and to well inform meetings participants before the face-to-face meeting takes place. In this sense, a data model to support group discussions in a pre-meeting context assuming an asynchronous and geographically distributed interaction was proposed.

Main results
- a conceptual argumentation model to support pre-meeting discussion
- prototypes implemented by partners in diverse platforms
- the creation of the International Workshop on Groupware (CRIWG) that will be in its 14th edition in 2008

Number of involved students

MSc	4
Undergraduate	2

Publications

Journals: 1
International Conferences and workshops: 5
National Conferences: 5

5. Project Name GeoTec

Title Geoinformatics: Methods and Techniques

Financial support CNPq/Protem-CC

Period 1995-1997

Partners Universidade Estadual de Campinas (Unicamp) - coordinator
Instituto Nacional de Pesquisas Espaciais (INPE)
Pontificia Universidade Catolica do Rio (PUC-Rio)
Universidade Federal de Pernambuco
Universidade Federal de Goiás (UFG)

Role Member

Abstract The geographical applications appeared as an interesting domain area. The main idea of this collaborative project was to use the acquired knowledge of all involved groups in spatial data modeling and querying to propose new methods and techniques.

Main results

- an object-oriented approach to model spatial data
- a query reformulation process
- specification and implementation of schema quality criteria
- the creation of the GeoInfo (Brazilian Symposium on Geoinformatics) that will be in its 10^{th} edition in 2008

Number of involved students

PhD	1
MSc	3
Undergraduate	2

Publications

International Conferences and workshops: 1
National Conferences: 7

ANNEXE 2

Quantitative Summary

This is a quantitative summary of the results achieved (concluded and ongoing PhD and MSc theses, publications, projects with financial support) in all research areas I have been working.

Geographical Databases
- PhD theses: 3
- MSc theses: 8
- Journals: 1 (submitted)
- International Conferences and Workshops: 6
- National Conferences: 18
- Projects: 4

Data Integration
- PhD theses: 1
- MSc theses: 5
- Journals: 1 + 2 submitted
- International Conferences and Workshops: 9
- National Conferences: 3
- Projects: 2

Semantic Issues in PDMS
- PhD theses: 2 concluded and 3 ongoing
- MSc theses: 1
- Journals: 2 + 1 (submitted)
- International Conferences and Workshops: 8
- National Conferences: 3
- Projects: 2

Cooperative Systems
- MSc theses: 4
- Journals: 1
- International Conferences and Workshops: 5
- National Conferences: 5
- Projects: 1 (international)

Information Retrieval
- MSc theses: 5
- Journals: 1
- International Conferences and Workshops: 4
- National Conferences: 1
- Projects: 1

Multimedia Databases
- MSc theses: 3
- Journals: 1
- International Conferences and Workshops: 4
- National Conferences: 8
- Projects: 1