ISSN (e): 2250 – 3005 || *Vol, 05* || *Issue, 01* || *January – 2015* ||
International Journal of Computational Engineering Research (IJCER)

Data Integration in Multi-sources Information Systems

Adham mohsin saeed
Computer Engineer Dept, Al-Rafidain University College,Iraq,Bagdad,

ABSTRACT

Data integration involves combining data coming from different sources and providing users with a unified view of these data. In this paper, we introduced the problem of integration of multi-source information systems with some focus on the heterogeneity and conflict issues at different levels .We have made some survey related the current implementations methods that have been used to solve the problems of data integration of multi sources IS which can be classified on three will established approaches, we have also discussed some of the limitations and advantages of such approaches, next we talked about the current trends in data integration such as warehousing, descriptive logic and ontology . Finally, we have presented some case studies that have been implemented using some of these approaches.

Key words: Data integration, multi-source information systems, warehousing.

I. INTRODUCTION

In recent years there is a tremendous growth for the need of a various application that can access, relate, use, and integrate of multiple disparate information sources and repositories including databases, knowledge bases, file systems, digital libraries, information retrievals systems and electronic mail systems. Data integration is a core issue in these applications, For instance, in the area of business intelligence (BI), integrated information can be used for querying and reporting on business activities, for statistical analysis, online analytical processing (OLAP), and data mining in order to enable forecasting, decision making, enterprise-wide planning, and in the end to gain sustainable competitive advantages. (1) A study by Forrest Research reported that 98% of companies it recently interviewed said that integration is either “extremely important” or “very important” to their firm’s IT strategy and their integration projects have been running for an average of more than 20 months and involve an average of seven systems . (2) Ensuring Data integrity in the process of integration of heterogeneous data sources has proven to be very challenging task because of the “asymmetric “nature of the integration and a “seamless Integration” has been so far more of a wish than reality, due to the difficulties and challenges faced in dealing with the heterogeneity, technology obsolescence and semantic discrepancy associated with multiple source information systems. (3) More recently, research has focused on the Intelligent Integration of Information. Here the problem is to access the diverse data residing in multiple, autonomous, heterogeneous information sources, and to integrate, or fuse, that data into coherent information that can be used by decision makers. (4)

II. MULTI-SOURCES INFORMATION SYSTEM

A **Multi-Sources Information System** (MSIS) can be defined as an Information System where exist a set of different User Views and a set of distributed, heterogeneous and autonomous Information Sources (5). We briefly explain these concepts: (6) (7)

- **Distributed**: Nowadays most computers are connected to some type of network, especially the Internet, and it is natural to think of combining application and data sources that are physically located on different hosts, but that can communicate through the network.
- **Heterogeneous** information sources: are sources with possibly different data models, schemas, data representations, and interfaces.
- **Autonomous** information sources: are sources that were developed independently of each other, and are maintained by different organizations, that may wish to retain control over their sources.

Figure (1) shows the architecture of such a system. There are three layers: source, mediation and application. The source layer contains each source with its associated wrapper, which translates queries and queries’ responses that pass through it. The mediation layer has in charge the transformation and integration of the information obtained from the sources, according to the requirements coming from the application layer. The application layer provides the user views to the user applications through execution of queries over the mediation layer. (5)

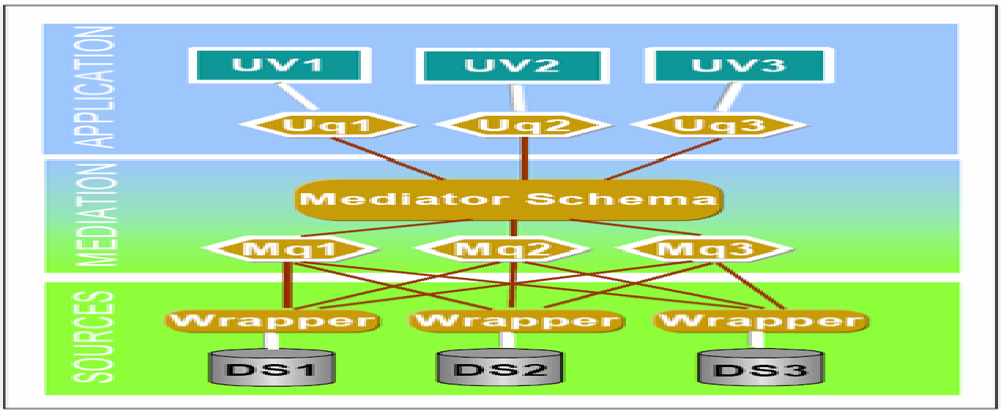

Figure (1) MSIS architecture

Some of the requirements that may be faced when designing a multi-source information system:

1. How to provide an integrated view of overlapping data sets from multiple sources.
2. How to support updates against such views.
3. How to identify the relationship between two or more institutions of replicated data.
4. How to keep replicated data “synchronized”.

III. INFORMATION CONFLICTS IN DATA SOURCES

With the advent of the internet and web technologies, the focus shifted from integrating purely well-structured data to also incorporating semi- and unstructured data while architecturally, loosely-coupled mediator and agent systems became popular. (1) (9)

An important issue almost in every system of information integration is the possibility of information conflicts among the different data sources. These conflicts are at two different levels: (6)

1. **Intentional inconsistencies**: The sources are in different data models, or have different schemas within the same data model, or their data is represented in different natural languages or different measurement systems. Such conflicts have often been termed **semantic inconsistencies**.
2. **Extensional inconsistencies**: There are factual discrepancies among the sources in data values that describe the same objects. Such conflicts are also referred to as data inconsistencies. Extensional inconsistencies can only be observed after intentional inconsistencies have been resolved. That is, different attribute names in the schemas of different information sources must be mapped to each other and attribute values must be within the same measurement system, to conclude that these values indeed contradict each other.

In a broad sense, semantic inconsistencies or conflict can occur at two different levels: at the **data level** and at the **schema level**. Data-level conflicts are differences in data domains caused by the multiple representations and interpretations of similar data. Data-level conflicts may have different forms such as data-value conflicts, data representation conflicts , data-unit conflicts , and data precision conflicts. (2)
Schema-level conflicts are characterized by differences in logical structures and/or inconsistencies in metadata (i.e., schemas) of the same application domain. Examples of such conflicts are naming conflicts, entity-identifier conflicts, schema-isomorphism conflicts, generalization conflicts, aggregation conflicts, and schematic discrepancies.

IV. APPROACHES TO INFORMATION INTEGRATION IN MULTI-SOURCE INFORMATION SYSTEMS

Information Integration can be defined as a process of using data abstraction to provide a single interface for viewing all the data within an organization, or a part of it, and a single set of structures and naming conventions to represent this data. (10) Data integration can be either virtual or materialized. In the first case, the integration system acts as an interface between the user and the sources, and is typical of multi databases, distributed databases, and more generally open systems. In virtual integration query answering is generally costly, because it requires accessing the sources. In the second case, the system maintains a replicated view of the data at the sources, and is typical, for example, both in information system re-engineering and data warehousing. In materialized data integration, query answering is generally more efficient, because it does not require accessing the sources, whereas maintaining the materialized views is costly, especially when the views must be up-to-date with respect to the updates at the sources (view refreshment). (11)

4.1 Established approaches to data Integration

Over the past few years various approaches and techniques have been proposed and adopted in the search for achieving an ultimate information integration system. In this section, we provide a brief description of the well-known approaches that form the basis for the design of existing integration framework.

Early work on integration was carried out in the context of database design, and focused on the so-called **schema integration** or **Mapping based approach** by constructing a global unified schema for semantically related information sources starting from several sub schemata (local schema), each one produced independently from the others. Mappings are not limited to schema components (i.e., entity classes, relationships, and attributes), but may be established between domains and schema components. (11) (2)

More recent efforts have been devoted to data integration, which generalizes schema integration by taking into account actual data in the integration process. Here the input is a collection of source data sets (each one constituted by a schema and actual data), and the goal is to provide an integrated and reconciled view of the data residing at the sources, without interfering with their autonomy. We only deal with the so-called read-only integration, which means that such a reconciled view is used for answering queries, and not for updating information. (11)

The second approach to the data integration problem is the **procedural approach** or **Intermediary approach,** where data are integrated in an ad-hoc manner with respect to a set of predefined information needs. In this case, the basic issue is to design suitable software modules (e.g., mediators, agents, ontology's) that access the sources in order to fulfill the predefined information requirements. Several data integration (both virtual and materialized) projects follow this idea. They do not require an explicit notion of integrated data schema, and rely on two kinds of software components: **wrappers** that encapsulate sources, converting the underlying data objects to a common data model, and **mediators** that obtain information from one or more wrappers or other mediators, refine this information by integrating and resolving conflicts among the pieces of information from the different sources, and provide the resulting information either to the user or to other mediators.. (11) (2).

The third approach is called **declarative approach** or **query-oriented approach**, here the goal is to model the data at the sources by means of a suitable language (either declarative logic-based languages or extended SQL) to construct a unified representation and to refer to such a representation when querying the global information system, and to derive the query answers by means of suitable mechanisms accessing the sources and/or the materialized views. This approach typically requires users to engage in the detection and resolution of semantic conflicts since it provides little or no support for identifying semantic conflicts (11) (2).

Note that research approaches classified into these three categories may not be mutually exclusive. For example, the intermediary-based approach may not necessarily be achieved only through intermediaries. Some approaches based on intermediaries also rely on mapping knowledge established between a common ontology and local schemas. It is also often the case that mapping and Intermediaries are involved in query-oriented approaches.

4.2 Current trend in data Integration

4.2.1 Warehousing

The Warehousing approach derives its basis from traditional data warehousing techniques. Data from heterogeneous distributed information sources is gathered, mapped to a common structure and stored in a centralized location. Warehousing emphasizes data translation, as opposed to query translation in mediator-based systems. In fact, warehousing requires that all the data loaded from the sources be converted through data mapping to a standard unique format before it is stored locally. In order to ensure that the information in the warehouse reflects the current contents of the individual sources, it is necessary to periodically update the warehouse. (12)

4.2.2 Description Logic

Description Logic provides a way to manipulate the semantics of data. It has the advantage that most of the data models developed so far can be represented using Description Logic. Systems based on Description Logics tend to focus on conjunctive queries i.e. these queries provide an integrated view, which hold a subset of data. This approach has been very successful in manipulating the meanings associated with data and the approach is also very flexible as the formalizations are based on concepts and roles. The basic essence of concepts and classes comes from Object Oriented (OO) Methodology. The strengths associated with OO approach can be utilized in this approach as well. This allows the developers to add new roles and reuse the already developed components. Some research material refers to these concepts as "classes". These concepts/classes provide the framework for managing meaning associated with data. (11)

4.2.3 Ontological

In the last decade, semantics (which are an important component for data integration) gained popularity leading to the inception of the celebrated ontology-based integration approach. The Semantic Web research community has focused extensively on the problem of semantic integration and the use of ontology's for blending heterogeneous schemas across multiple domains. Their pioneering efforts have provided a new dimension for researchers to investigate the challenges in information integration. A number of frameworks designed using ontology-based integration approaches have evolved in the past few years.

There are a lot of advantages in the use of ontologies for data integration. Some of them are : the ontology provides a rich, predefined vocabulary that serves as a stable conceptual interface to the databases and is independent of the database schemas; the knowledge represented by the ontology is sufficiently comprehensive to support translation of all the relevant information sources; the ontology supports consistent management and the recognition of inconsistent data. (7) (12)

V. CASE STUDIES

5.1 Infomaster: Mapping based approach

Infomaster is an information integration system that provides integrated access to multiple distributed heterogeneous information sources on the Internet, thus giving the illusion of a centralized, homogeneous information system. We can say that Infomaster creates a virtual data warehouse. Infomaster handles both structural and content translation to resolve deference's between multiple data sources and the multiple applications for the collected data. Infomaster connects to a variety of databases using wrappers, such as for Z39.50, SQL databases through ODBC, EDI transactions, and other World Wide Web sources. (13)

The core of Infomaster is a facilitator that determines which sources contain the information necessary to answer the query efficiently, designs a strategy for answering the query, and performs translations to convert source information to a common form or the form requested by the user. Formally, Infomaster contains a model-elimination resolution theorem prove as a workhorse in the planning process. There are wrappers for accessing information in a variety of sources. For SQL databases, there is a generic ODBC wrapper. There is also a wrapper for Z39.50 sources. For legacy sources and structured information available through the WWW, a custom wrapper is used.

Infomaster uses rules and constraints to describe information sources and translations among these sources. These rules and constraints are stored in knowledge. For efficient access, the rules and constraints are loaded into Epilog, a main memory database system from Epistemic.

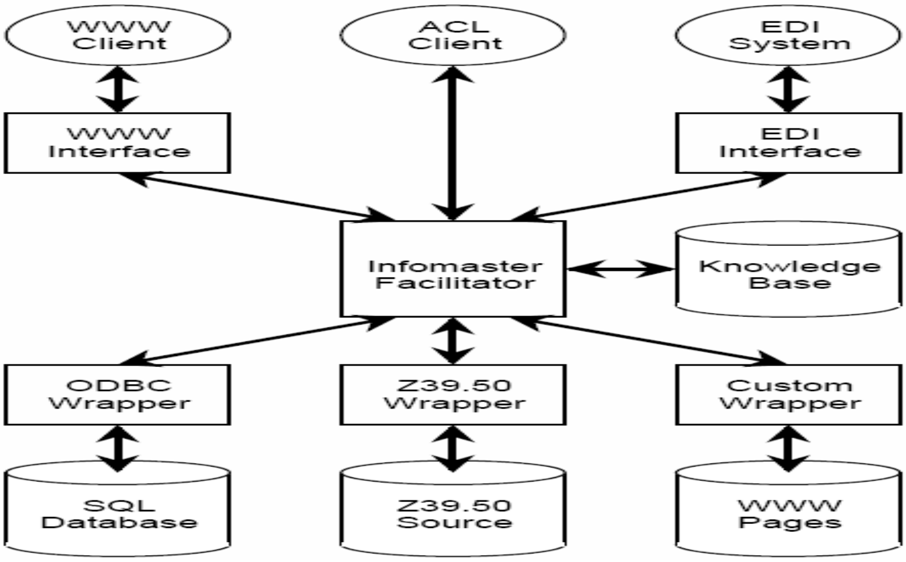

Figure 2: Infomaster Architecture

Infomaster includes a WWW interface for access through browsers. This user interface has two levels of access: an easy-to-use, forms-based interface and an advanced interface that supports arbitrary constraints applied to multiple information sources. However, additional user interfaces can be created without affecting the core of Infomaster. Infomaster has a programmatic interface called Magenta, which supports ACL (Agent Communication Language) access. ACL consists of KQML (Knowledge Query and Manipulation Language), KIF (Knowledge Interchange Format), as well as vocabularies of terms.

5.2 The TSIMMIS Project: an Intermediary approach to data integration

TSIMMIS is a mediator system being developed by the Stanford database group, in conjunction with IBM. The goal of the TSIMMIS project is to develop tools to rapidly accessing in an integrated fashion multiple information sources that may include both structured and semi-structured data and to ensure that the information obtained is consistent. (14)

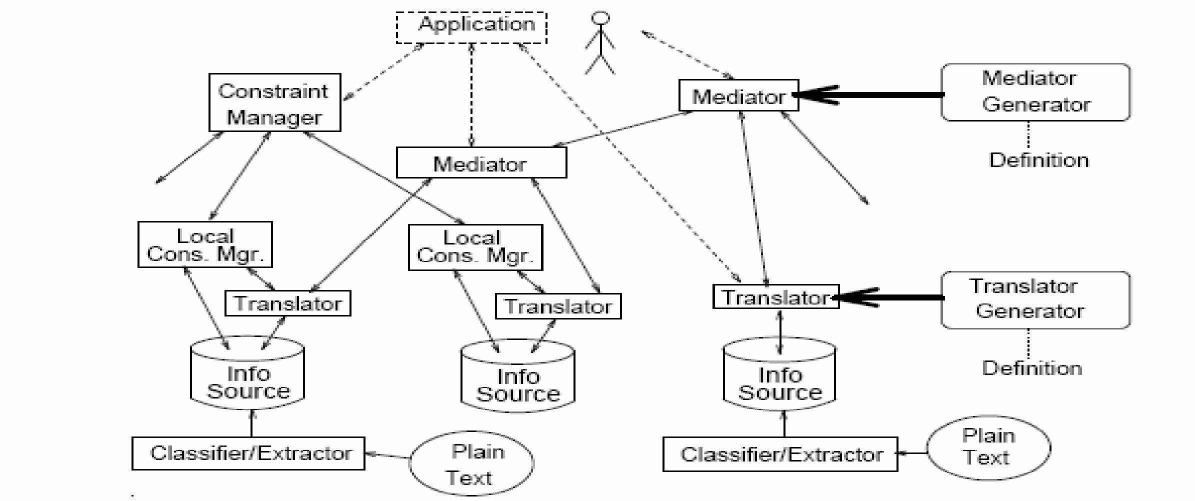

Figure 3: TSIMMIS Architecture

TSIMMIS employs classifiers/extractors, wrappers/translators, mediators, and constraint managers to achieve system integration. The above figure shows a collection of (disk-shaped) heterogeneous information sources. The classifiers/extractors attempt to identify simple patterns in unstructured sources and then export this information to the entire TSIMMIS system through a wrapper. Above each source is a translator (or wrapper) that logically converts the underlying data objects to a common information model. To do this logical translation, the translator converts queries over information in the common model into requests that the source can execute, and it converts the data returned by the source into the common model. For the TSIMMIS project they have adopted a simple self-describing (or tagged) object model called the Object Exchange Model, or OEM. One of the goals of this project was to automate the development of wrappers. To this end, a wrapper implementation toolkit was developed. The toolkit allows for the semi-automatic creation of wrappers through the use of predefined templates.

Above the translators in Figure 3 lie the mediators. A mediator is a system that refines in some way information from one or more sources. A mediator embeds the knowledge that is necessary for processing a specific type of information. The mediator may also process answers before forwarding them to the user, say by converting dates to a common format, or by eliminating articles that duplicate information.

Mediators export an interface to their clients that is identical to that of translators. End users can access information either by writing applications that request OEM objects, or by using one of the generic browsing tools that have been developed. Finally, there are the constraint managers which attempt to ensure semantic consistency across integrated resources.

5.3 Information Integration Wizard (IWIZ): Hybrid Approaches

Information Integration Wizard combines the data warehousing and mediation approaches. IWIZ allows end-users to issue queries based on a global schema to retrieve information from various sources without knowledge about their location, API, and data representation. However, unlike existing systems, queries that can be satisfied using the contents of the IWIZ warehouse are answered quickly and efficiently without connecting to the sources. In the case when the relevant data is not available in the warehouse or its contents are out-of-date, the query is submitted to the sources via the IWIZ mediator; the warehouse is also updated for future use. An additional advantage of IWIZ is that even though sources may be temporarily unavailable, IWIZ may still be able to answer queries as long as the information has been previously cached in the warehouse.

Due to the popularity of the Web and the fact that much of the interesting information is available in the form of Web pages, catalogs, or reports with mostly loose schemas and few constraints, IWIZ have focused on integrating semi structured data. The following figure show Schematic description of the IWIZ architecture and its main components. (8)

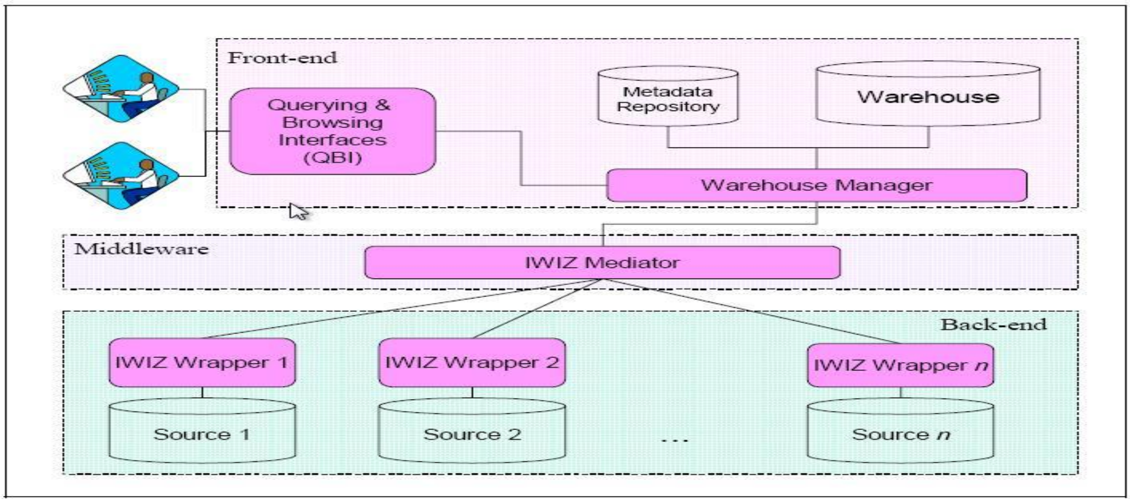

Figure 4: IWIZ Architecture

A conceptual overview of the IWIZ system is shown in the above Figure. System components can be grouped into two categories: Storage and control. Storage components include the sources, the warehouse, and the metadata repository. Control components include the querying and browsing interface (QBI), the warehouse manager, the mediator, and the wrappers. In addition, there is information not explicitly shown in the figure, which includes the global schema, the queries and the data. The global schema, which is created by a domain expert, describes the information available in the underlying sources and consists of a hierarchy of concepts and their definitions as well as the constraints. Internally, all data are represented in the form of XML documents, which are manipulated through queries expressed in XML-QL. The global schema, which describes the structure of all internal data, is represented as a Document Type Definition (DTD), a sample of which is shown later in the paper. The definition of concepts and terms used in the schema is stored in the global ontology.

As indicated in Figure 4, users interact with IWIZ through QBI, which provides a conceptual overview of the source contents in the form of the global IWIZ schema and shields users from the intricacies of XML and XML-QL. QBI translates user requests into equivalent XML-QL queries, which are submitted to the warehouse manager. In case when the query can be answered by the warehouse, the answer is returned to QBI. Otherwise, the query is processed by the mediator, which retrieves the requested information from the relevant sources through the wrappers. The contents of the warehouse are updated whenever a query cannot be satisfied or whenever existing content has become stale. Our update policy assures that over time the warehouse contains as much of the result set as possible to answer the majority of the frequently asked queries.

VI. CONCLUSIONS

Integration of multiple information sources aims at combining selected sources so that they form a unified new whole and give users the illusion of interacting with one single information sources. In this paper, we briefly introduced the problem of integration of multi-source information systems with some focus on the heterogeneity and conflict issues at different levels that may affect such a design of such systems and should be considered when designing such systems. We have made some survey related the current implementations methods that have been used to solve the problems of data integration of multi sources IS which can be classified on three will established approaches, we have also discussed some of the limitations and advantages of such approaches, next we talked about the current trends in data integration such as warehousing, descriptive logic and ontologies. Last, but not least, we have presented some case studies that have been implemented using some of these approaches.

REFERENCES

[1]. **Patrick Ziegler, Klaus R. Dittrich.** *Three Decades of Data Integration - All Problems Solved ,2004.* s.l. : In 18th IFIP World Computer Congress, 2004. Vols. Volume 12, Building the Information Society , 2004.

[2]. **PARK, JINSOO and RAM, SUDHA.** *Information Systems Interoperability:What Lies Beneath?* s.l. : ACM Transactions on Information Systems, Vol. 22, No. 4, 2004.

[3]. **Y. Tang, J. B. Zhang, C. H. Tan and M. M. Wong.** *A Five-step Approach to Multi-source Information Infusion with Guaranteed Data Integrity.* s.l. : Singapore Institute of Manufacturing Technology (SIMTech) web site , 2003.

[4]. **Weishar, Larry Kerschberg and Doyle J.** *Conceptual Models and Architectures for Advanced Information Systems.* s.l. : Applied Intelligence, vol. 13, pp. 149-164, 2000.

[5]. **Marotta, Adriana.** *Quality Management in Multi-Source Information Systems.* s.l. : Instituto de Computación. Facultad de Ingeniería.Universidad de la República. Montevideo, Uruguay., 2004.

[6]. **Motro, Amihai and Anokhin, Philipp.** *Fusionplex: resolution of data inconsistencies in the integration.* Department of Information and Software Engineering, George Mason University, University Drive, Fairfax, VA 22030-4444, USA : Information Fusion 7 176–196, 2006.

[7]. **Agustina Buccella, Ra Cechich, Nieves R. Brisaboa.** *An Ontology Approach to Data Integration.* s.l. : Journal of Computer Science and Technology , Vol 3,No. 2, 2003.

[8]. **Hammer, Joachim and Pluempitiwiriyawej, Charnyote.** *Overview of the Integration Wizard Project for Querying and Managing Semistructured Data in Heterogeneous Sources.* s.l. : Proceedings of the Fifth National Computer Science and Engineering Conference , Thailand, , November 2001.

[9]. **Wood, Peter.** *Semi-Structured Data.* Birkbeck College at the University of London : http://www.dcs.bbk.ac.uk/~ptw/, 2003.

[10]. **Michiels, Eric.** *New Trends in Information Integration.* s.l. : IBM corporation ,http://www.econ.kuleuven.be/, 2008.

[11]. **Calvanese, Diego.** *Description Logic Framework for Information Integration.* s.l. : Proc. of the 6th Int. Conf. on the Principles of Knowledge Representation and Reasoning , 1998.

[12]. **Chakravarthy, Aditya Telang and Sharma.** *Information Integration across Heterogeneous Domains: Current Scenario, Challenges and the InfoMosaic Approach.* s.l. : Department of Computer Science and Engineering, University of Texas at Arlington, 2007.

[13]. **Genesereth, Michael R., Keller, Arthur M. and Duschk, Oliver M.** *Infomaster:An Information Integration System.* s.l. : proceedings of 1997 ACM SIGMOD Conference , 1997.

[14]. **Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., and Widom j.** *The TSIMMIS project: Integration of heterogeneous information sources.* Tokyo : 10th Anniversary Meeting of the Information Processing Society of Japan, 1994.