Information integration among Heterogeneous and Autonomous Applications

Benabdelkader, A.

Publication date
2002

Citation for published version (APA):
Benabdelkader, A. (2002). *Information integration among Heterogeneous and Autonomous Applications*. Febo Druk.

Chapter 1

Introduction

The design and development processes of advanced applications in scientific and system engineering domains consider different data modeling and information management strategies. Data models define the data structures and relationships among the data, to reflect the proper representation of the information each application needs. The information management strategies however, depend on the global architecture design and the chosen database system to fulfill the functionalities required by the application.

Diversity of the used information management approaches is usually due to different characteristics and requirements of each application. Due to the complex requirements of emerging applications, several scientific and business oriented organizations from biology, medicine, physics, astronomy, engineering, e-commerce, etc. have realized the need to reconsider their information management systems towards better addressing of collaborative work. Therefore, these organizations are required to provide appropriate products and services, and to better react to the new information management requirements in terms of data integration. Traditionally, information integration and data translation among different heterogeneous and autonomous sites were considered a completely manual process, where either the user or the database administrator must do the data translation and exchange. Nowadays, the problem of data integration and information exchange among heterogeneous data sources has become a challenging issue to be studied, and different integration approaches are being examined and evaluated.

This thesis addresses the issue of *information integration* for *systems interoperation* among heterogeneous and autonomous applications, and mainly addresses solutions related to the requirements for:

- *Data integration* from different sources distributed over a network of nodes.
- *Interoperation* and *information exchange* among a number of sites, which are heterogeneous, autonomous, and of distributed nature.
- *Methodology* design and *generic tools* development to support the information integration amongst a number of networked applications.

Within the different chapters of the thesis we propose *methodologies*, develop *standard tools for information access/exchange*, and validate *generic solutions* that serve the information management requirements. Generic solutions fit several applications emerging from various domains, and facilitate systems flexibility and configurability. The design and development of generic solutions, for information management and interoperation, is achieved via the deployment of standards and middleware solutions during the different development phases of an application, which evolve from modeling and design, to development and validation. In addition, the proposed solutions for information integration and systems interoperation consider the combination of emerging advances in databases and Web technologies, and mainly deploy the related standard concepts for data modeling, data definition, information storage and retrieval, information exchange, multi-platform-programming environment, and Communication infrastructure. Considering the main characteristics defining advanced applications, described above, the structure of the thesis is motivated by the following facts:

① There is wide variety of emerging networked applications in the diverse domains of science, business, engineering, education, e-commerce, tourism, etc. Among the common characteristics of these applications, we can enumerate distribution diversity, site autonomy, and information heterogeneity. In addition, the use of different data modeling approaches, information management strategies, and database management systems complicate the interoperation among these applications.

② Based on the type of applications and the global objectives targeted by each of them, the information management requirements differ from one application to another.

③ To fulfill the information management requirements of these applications, several approaches have been proposed and developed, addressing the information integration problem. Most of these approaches are application specific, while attempts in the direction of using standard tools and generic solutions are quite a few.

④ In order to validate different approaches addressing specific requirements of these applications, it is necessary to design and develop some prototypes as proof of concepts in different applications.

⑤ Based on the studies and prototypes developed within the different projects addressed in this thesis, a 'challenging' generic and flexible approach is designed and presented. This approach benefits from previous methodologies and extends them in order to provide generic solutions that can be applied within a wide variety of applications.

The five major points enumerated above illustrate the reason for the structure of the thesis. The application cases, mentioned in point 4 above and addressed in chapters 3, 4, and 5 of this dissertation, provide the base for the Generic and Flexible Information Integration System (**GFI$_2$S**) addressed in chapter 6. Mainly, the **GFI$_2$S** system reflects the results of the lessons learned in three research projects, and thus, addresses the integration of data sources, and the interoperation among diverse, heterogeneous, and autonomous sites in a network.

This introductory chapter briefly emphasizes the main characteristics of emerging applications and outlines the major requirements of these applications in terms of information management and interoperation. Chapter 2 describes some of the existing related approaches, mechanisms, and tools for information integration. Three specific research projects from different application domains are described in chapters 3, 4, and 5 of this thesis. These chapters focus mostly on the study and analysis of the information management requirements for distributed and heterogeneous applications, as inspired by the advanced cooperative applications in the domains of water systems industry, e-commerce, and e-science. The E-Science[1] domain addressed in chapter 5 mostly focuses on how to

[1] E-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it. *John Taylor, Director General of the Research Councils, OST.*

use scientific methods, and how to apply new software packages and web resources in the analysis and solutions of real application problems, and in specific for the experiments in research laboratories. Chapter 6 of the thesis addresses an open and flexible approach for information integration and systems interoperation (**GFI$_2$S**). First it outlines the main requirements for emerging applications, and then it defines a common integrated approach that can support many emerging applications from different application domains. The approach also applies object-oriented standards and middleware solutions in order to provide more generic utilities.

1.1 Major Requirements in terms of Information Management

This section outlines the major requirements for future advanced applications in terms of information management and systems interoperation. These requirements are mainly identified through the three application cases addressed in the thesis, in addition to some other research work and literature. A detailed description of these requirements is out of the scope for this introductory chapter, appropriate descriptions however, are given in different chapters of this dissertation. It is also necessary to mention that these requirements are identified to support the information sharing and the data integration among networked applications, while preserving their local autonomy, heterogeneity, and distribution.

Hereafter, we enumerate a list of the major requirements that need to be addressed when designing and developing appropriate information management strategies for advanced applications. In order to give a better understanding and a clear overview to the reader, these requirements are grouped into six main categories:

R$_1$: Information Integration for Systems interoperation

- Transparent access to data located at different sources, via on-line integrated views.
- Interoperation among different systems in term of information exchange and services.
- Information sharing within a large community of internal and external applications and users.
- Data integration from heterogeneous and distributed sources.

R$_2$: Security for access and visibility levels

- User authentication based on pre-defined access rights.
- Information visibility levels based on pre-defined import/export schemas.
- Separation between public and proprietary information.

R$_3$: User facilities

- Provide easy access to data independent of its internal format and structure.
- Support user friendly interfaces facilitating the exploration of information, characterized by its complex structure.

R_4: Use of standards and middleware solutions

- Universal accesses to data regardless the underlying database management system.
- Use of standards and middleware solutions for data modeling and information exchange.
- Support for multi-platform applications development.
- Scientific data classification and cataloguing via the deployment of emerging standards in the field such as Dublin Core for scientific data description and NetCDF[2] for array-oriented data representation.

R_5: System efficiency and effort Minimization

- System efficiency and performance for data manipulation.
- Short response time for on-line requests.
- High bandwidth for data transfer.
- Good strategies for data storage and data archives.
- Effort and cost minimization in both modeling and development phases.

R_6: Advanced features

- Support for new data types introduced within the scientific application domain.
- Support the management of large data sets.
- Combine databases and advanced web technologies.
- Combine object-oriented concepts and emerging standards.

1.2 Application Cases: an Overview

The application cases presented in chapters 3, 4, and 5 of this dissertation illustrate three examples of modern applications. These application cases are primarily addressed by their data modeling concepts and information management strategies, that are required to support each application domain. The choice of these three application cases have covered different domain criteria on data integration mechanisms. In addition, these applications constitute some of the research areas within the CO-IM[3] group at the University of Amsterdam, targeting the COoperative Information Management among autonomous and heterogeneous applications. Furthermore, the diversity of these application domains have added more value to our research work by introducing various requirements and thus, adding new challenges. The three applications addressed in this dissertation include:

- Intelligent supervision and control in heterogeneous and distributed water environments (Waternet project).
- Interoperation and collaboration among large distributed databases for music industry and e-commerce applications (MegaStore project), and
- Scientific data archiving and cataloguing for e-science within the Virtual Laboratory environment (Virtual Laboratory project).

[2] Network Common Data Form

[3] CO-IM: COoperative Information Management group of the University of Amsterdam (`http://www.science.uva.nl/~netpeer`)

The study of these applications shows that even while addressing only the information management issue, the requirements are quite complex and specific for each application domain. Thus, the modeling constructs, the designed methodologies, and the used systems differ from one application to another. More precisely:

1. In water supply industries (Waternet) distinct functionalities required in this industry are supported by independent, heterogeneous, and autonomous subsystems. Each subsystem performs its specific activity, but their co-working and complex information exchange needs to be properly supported in order to assure a continuous supply, to meet the quality standards, to save energy, to optimize pipeline sizes, and to reduce wastes.

2. In music industry application (MegaStore), we address the design and development of advanced and efficient internet-based Electronic Commerce (E-Commerce) services to support necessary requirements for the buyers of different goods. In addition to the traditional user requirements for every application environment, the developed system properly addresses several efficiency, organization, and multimedia related issues. Among the addressed issues, we enumerate: the data catalogues and information classification, short response time for on-line requests, high system performance, and high data transfer rates.

3. In experimental life science application (Virtual Laboratory), the information management framework aims at developing digital libraries and toolkits to enable scientists and engineers to work on their problems via experimentation. Especially, the VL information management framework addresses some emerging issues related to the management of large multimedia scientific data, information integration from a variety of data sources, and collaborative developments in e-science environment.

As depicted in Table 1.1, a large list of requirements is addressed in each of these research projects. On one hand the three projects are characterized by needs from different application domains, thus the requirements differ slightly from one project to another. On the other hand, considering the fact that the projects are carried out in different periods also shows that some requirements change in time, to cope with the level of advances in different applications areas. These application cases are not to be directly compared with each other, rather they, in one way or another, complement each other in creating a more comprehensive set of requirements and contribute to the design and development of the (**GFI$_2$S**) solution. In each of the application cases, some of the requirements do not apply due to the type of application (symbolized by Φ); and some others were not addressed due the main aim of the application (symbolized by X). As such, the Waternet is a specific peer-to-peer system in which, the use of standards and middleware was not obligatory, while the MegaStore project presents a system for a large community in which, the use of common technologies in databases and Internet is mandatory.

Even at the level of each individual requirement, its consideration may be partially addressed within different research projects. For instance, the requirement of *the use of standards and middleware solutions* represents several concepts, which are partially addressed within the three projects. As such, the MegaStore system addresses the standards at the level of data modeling (e.g. UML), data definition (e.g. ODL), and information storage/retrieval (e.g. SQL/SQL3, OQL). While, the Virtual laboratory information management framework extends the use of standards to also support universal data access and scientific data modeling. Similarly, the data integration and the information sharing are addressed at different levels of complexity within each project.

Major Requirements in terms of Information Management	**Waternet 1996-1998**	***MegaStore 1999-2000***	**Virtual Lab 1999-2003**	**GFI$_2$S**
Information integration for systems interoperation				
- Transparent access to data	√	√	√	√
- Systems Interoperation	√	Φ	√	√
- Information sharing	√	Φ	√	√
- Data integration	√	Φ	√	√
Security for access and visibility levels				
- User authentication	√	√	√	√
- Information visibility levels	√	Φ	√	√
- Separation between public and private data	Φ	√	√	√
User Facilities				
- Easy access to data	X	√	√	√
- User friendly interface	X	√	√	√
Use of standards and middleware				
- Universal access to data	X	Φ	√	√
- Standard Data modeling	Φ	√	√	√
- Multi-platform development	X	X	√	√
- Data classification and catalogs	Φ	√	√	√
System efficiency and minimization				
- System efficiency and performance	X	√	√	√
- Short response time for user requests	X	√	√	√
- High bandwidth for data transfer	X	√	√	√
- Good strategies for data storage	X	X	√	√
- Effort and cost Minimization	X	√	√	√
Advanced Features				
- Support for new data types	Φ	√	√	√
- Management of Large data sets	Φ	√	√	√
- Combine advanced databases and web technologies	Φ	√	√	√
- Combine O-O concepts and emerging standards	X	X	√	√
Notation: √ :Addressed | Φ: Does not apply | X: Not addressed				

Table 1.1: List of Requirements of Today's and Forthcoming Applications

As depicted in column 4 of Table 1.1, the richness of the VL advanced applications, which emerge from various scientific domains, requires the consideration of most defined requirements. In addition, the Generic and Flexible Integration System (**GFI$_2$S**) presented in chapter 6 also considers the totality of these requirements in order to achieve a more flexible and open solution, serving the interoperation among advanced applications.

The list of requirements presented in Table 1.1 can also be categorized into two main categories of:

- Several information *Research Challenges* that need to be addressed in order to support the integration of information among networked applications, while preserving their local autonomy, heterogeneity, and distribution. Some of the *Research Challenges* that are addressed in different chapters of the thesis document include: information integration, security for access and visibility levels, and other advanced features such as supporting new data types and applying standards to object-oriented concepts.
- Several *standard tools* and *advanced Web technologies* that need to be applied for information management, to facilitate the information exchange among interoperable systems. Namely, standard data access, user friendly interfaces, and support for new data types.

The contribution of this thesis to the area of information integration and system interoperation is to investigate some of the research challenges and to apply emerging technologies and database standards to data integration mechanisms. Previous work on information integration for systems interoperation has developed some specific solutions for information exchange and data integration. These solutions are characterized by their specific tools and languages, which are mostly difficult to learn and hard to maintain. Our approach however, applies emerging standards and Internet technologies for information integration among autonomous and heterogeneous sites. This approach presents an open information integration facility for heterogeneous systems, while preserving their autonomy and distribution.

1.3 Thesis Contribution

The emerging advanced applications from scientific and business organizations present new challenges to the research in the domain of information management and interoperation. The challenges, primarily include: handling multi-media data types from the scientific domain, and deploying new co-working environments mainly based on the Internet technology, middleware, and standard solutions. MiddleWare in this thesis document is a general term representing any developed software that serves to "glue together" or mediate between two separate and usually pre-existing programs. For instance, a common application of middleware is to support programs written at one site with a particular database for access to other databases. As standard solutions to be applied to the information integration for systems interoperation, this thesis addresses: ODL for data definition, OIF/XML for objects representation, ODBC/JDBC for database connection, and XML as a flexible way to create common information format and share the format together with the data on the World Wide Web, intranets, and elsewhere.

Research in the domain of information exchange, interoperability[4], and data source integration is still an open area for advanced architectures design, integration mechanisms, and tools development. Furthermore, the new emerging Internet technology for applications communications and middleware solutions for universal data access via standards offer promising approaches for the research in this area. Thus, there is a growing need for new

[4] Interoperability is basically the ability of a system or a product to work with other systems or products without special and extra effort on the part of the end-user

approaches to be designed and tools to be developed, in order to support interoperable information systems and to facilitate their collaboration mechanisms.

The work presented within this dissertation is an *"application driven database research"*, to better support the real information management needs and requirements of advanced emerging applications. It also introduces new approaches to integrate different types of data from heterogeneous applications to achieve broader information access, minimizing specialized development efforts, and attain competitive advantages.

In order to handle the new emerging data types, the work described in this dissertation document addresses two complementary trends. The first is to extend the traditional database systems to handle new and different data types and migrate these types of data into the DBMS. The second is to apply the middleware approach to provide standardized interfaces for all types of data, maximizing interoperability and reusability, while leaving the data in the place where it is generated or heavily used.

The main aim of the thesis is to address the design and partial development of a generic system to support information sharing among a wide variety of applications, and to assist their proper collaborative working environment, and flexible information integration.

Based on the expertise gained in the design and development of the various R&D projects during this Ph.D. study and based on the investigation, evaluation, and validation of the methodologies and systems discussed in chapters 3, 4, and 5 of the dissertation, an open and flexible integration approach is presented in chapter 6. The flexibility of this approach is achieved via some database extensions and through the deployment of object-oriented standards, emerging Internet technologies, and middleware solutions. The database modeling and the scientific data cataloguing and archiving, are supported via the deployment of the object-oriented standards. While, the emerging Internet technologies and middleware solutions allow universal access to the data, ease the information exchange mechanisms, and facilitate the multi-platform applications development. Therefore, *the thesis contribution to the research area of information integration resides in the specific combination of emerging standards in the filed with the fundamental research approaches and the way in which they are inter-linked.*

The approach we propose does not only take into consideration the work that has been done in the area of information exchange and interoperation. But, it also considers the utilization of both advanced web and database technologies in order to support the new requirements from the scientific applications for handling large multimedia data sets, and applies the new emerging technology for Internet communications and universal data access through middleware and standard solutions. As such, the developed framework and the designed methodology enable easy cooperative work among existing systems, while preserving their autonomy and heterogeneity.

1.4 Organization of the thesis

The remaining of this dissertation document is organized as follow:

- In Chapter 2, we study, analyze, and discus a number of classifications for information management and the state of the art in data integration approaches. This study considers both distributed and integrated systems, and illustrates the need for these distinct approaches in order to support the complex requirements of different advanced cooperative applications from system engineering to scientific domains.

- Chapters 3 and 4 present two case studies in the field of intelligent supervision/control in heterogeneous and distributed applications. namely. the water distribution management and the advanced word wide databases for e-commerce. The designed and developed frameworks for these specific cases will be evaluated. This evaluation allows the validation of the most important and relevant features. that need to be taken into consideration when designing the flexible integration approach. These beneficial features are further addressed and deployed within the integration approach presented in chapter 6.

- Chapter 5 addresses a third application case from the scientific domain; namely, the information management framework of the Virtual Laboratory project. The Virtual Laboratory project addresses the issue of manipulating large scientific and engineering data sets in terms of data acquisition from heterogeneous resources. information modeling of complex and varied application types from several large scientific emerging domains, data archiving mechanisms for very large objects (e.g. binary and text), and resources cataloguing based on using the Dublin Core standard model. The information management approach of VL. focuses on the use of middleware and standard *de facto* solutions as a means to enforce and standardize the information access/retrieval processes among multidisciplinary applications. The proposed approach incorporates several advanced key features to support system efficiency and performance, enforced by the provision of security for access, and visibility rights to information, database indexing, cataloguing mechanisms, and database performance analysis.

- Chapter 6 proposes a Generic and Flexible Information Integration System (**GFI$_2$S**) based on the investigation. evaluation. and validation of the methodologies and systems discussed within the previous chapters. The extensibility of **GFI$_2$S** is achieved via several data management functionality extensions and through the deployment of object-oriented standards. emerging Internet technologies. and middleware solutions. An important distinction between the system we are designing and other integration systems resides in the introduction of the two components in the architecture of **GFI$_2$S**. The first component of the architecture (called *Local Adaptation Layer*) assures proper communication between the local data source and its federated layer. While. the second component (called *Node Federation Layer*) presents the node's window to the outside world for information sharing and interoperation.

- Finally, chapter 7 concludes the thesis and summarizes the main conclusions derived from this research and provides examples of the **GFI$_2$S** deployment in real applications.