ARCHITECTURE FOR HETEROGENEOUS INFORMATION INTEGRATION PROCESS IN AN ECONOMIC INTELLIGENCE SYSTEM

Gbola IDOWU, Odile THIERY
idowugbo@loria.fr, Odile.Thiery@loria.fr
SITE-LORIA Nancy Université, B.P. 239 54506 Vandoeuvre, France

Mots clefs :
Métas données, système d'informations stratégiques, entrepôt documentaire, Méta modélisation.

Keywords:
Meta data, strategic information system, data warehouse, Meta modelling.

Palabras clave :
Meta datos, sistemas de información estratégica, almacén de datos, modelado de Meta.

Abstract

As economic intelligent process utilizes information obtained from disparate and heterogeneous information sources. We propose an architecture to integrate the information required to support a decision making process.
The architecture relies on the objective of the user and the well expressiveness of the user's information need in its information integration efforts. We try to exploit the information need of the user in order to map relevant sources which could meet the description of his information needs. Therefore, complementing a user profile with the information needs to enhance a user model that could have been conceptualized according to the user' role within an enterprise.

1 Introduction

Heterogeneous information integration is an attempt to provide a common information access interface to several sources of information whose contents, nature and accessibility are different. There is difficulty in providing access to a large collection of information sources since several sources store inter related information, and any query answering system must understand and exploit the relationships between their contents. This could become more difficult especially when the nature of their contents is different in respect to data encoding and, storage technology and presentation [7]. To enhance the approaches of heterogeneous data integration efforts in [1, 2, 3, 4, 12, 28], we attempt to develop an architecture for a materialized heterogeneous information integration system for an Economic Intelligent System that can satisfy an information access requirements for a system of strategic information defined in [1]. However, an information accessing system that would meet a universal usability requirement should accommodate a diverse set of users [26]. A user can have diverse sets of information requirements for a decision support purpose and indeed, depicts diverse behavior towards information available or obtained during his information search activities.

The growth of information technology, the World Wide Web and computer use in general, users with a variety of backgrounds, skills, interests, preferences, expertise and learning styles uses computer for purposes ranging from home entertainment to collaborative, mission critical processes like decision support. Therefore there are still many challenges to be met before a global universal usability can be satisfied. No single computer -user interaction have satisfied all users. Users have different needs and preferences to how their interaction with the system and the presentation of the result of their information search.

An attempt has been made in [2] to model an adaptive information retrieval system by user modeling approach. As extension to this effort of user modelling, we make an effort to propose architecture for an Economic Intelligent system that will take into consideration the user model in an attempt to integrate information from disparate information sources into a coherent store for decision support purpose in an Economic intelligent process.

In section 2 of this paper, we discussed the need for a data warehouse in decision support activities as related to Economic intelligent process. Section 3 explains the relevance of the user profile in pursuit of information search to meet an information need while architecture for an information integration process in EI was introduced in section 4 and we conclude this paper in section 5.

2 Economic Intelligence and Decision Support Systems.

Economic Intelligence (EI) has been defined as “the use of information in decision-making” – [14]. The EI process starts with the identification of a decision problem which is transformed into information need(s) [6]. There is a need of an individual who would be charged with the responsibility of looking for information to solve the decisional problem [7], after a rigorous process of the transformation of the decision problem into an information search problem [16]. He interacts with the information base of the organization which is fed from disparate operational sources. These operational sources could be among the online transactional processing systems (OLTP) of the organization which are used primarily for the transactional activities of an enterprise. These sources are referred to as internal information sources since the

administrators, authors and the contents are derived from the activities of the organization. Otherwise they are external sources to the organization concerned with the use of the information.
It has been observed by W. Inmon [17] and E. F. Codd [11] in the 1990s that the OLTP and the decision support application (OLAP – online analytical processing system) cannot coexist efficiently in the same database environment, mostly due to their very different transactional characteristics. Therefore, data warehousing took a much broader role, especially in the provision of information base to the OLAP systems. Here, data warehousing is seen as a strategy to bring heterogeneous information together under a common conceptual and technical umbrella and to make them available for decision support since most decision supporting systems require historical data in order to analyze trends and forecast what the data means to the existence of the organization.
We are of the opinion that, there is a need to understand the information needs of the end users in order to determine the structure of the strategic information system. In EI, a data warehouse becomes strategic to an enterprise if it can support strategic decision making process of the enterprise. Also, if the goal of a SIS is to supply the right information to the right users at the right time then the SIS could function as an EI system. It is important to note that response time to an information search in EI process should be a deterministic factor for a system to qualify as an EI system.

3 User profile and Enterprise Decision Support System

In SITE-LORIA, we conceptualized an user model in respect to information search and retrieval to contain: general knowledge of seeking behavior, user preferences, values and expectations, user intentions in relation to decisional problems in Economic Intelligence, as well as user knowledge status and level [2]. The above concept is contained in the user's responsibility or function in an Economic Intelligence process.

We have been able to distinguish several actor types in Economic Intelligence such as the decision maker, the watcher, the organization partners, and members of staff [19]. A decision-maker has been defined [13] as "The person that is capable of identifying the problem to be solved in terms of stake, of risk, and/or of threat which weighs on the enterprise". The watcher is the information specialist; he is best placed to gain intelligence about networks, databases and contexts of the enterprise. He is a fundamental intermediary between the decision-maker and the information base. The model of the decision maker and the watcher are described in [6].

Here, we to focus more on the object of search which could be as a result of the decisional problem at hand and the profile of actors. Therefore, we could attach to each user its own peculiar information needs in order to retrieve and integrate information according to his perspective of the solution to the problem that arises. We could achieve this by mapping a user profile to information sources that are relevant to the informational need of the user and therefore, retrieve such information to be stored in a repository (Data Warehouse). This can allow us reuse information retrieved in the process of trying to resolve a certain decisional problem if a similar problem arise in future.

Also, we should be able to associate derived views of stored information in the repository as a formulated query from the user' information need. The later could be referred to as a personalized data mart, that is, a derived mart based on a user profile. Queries could be made across several data marts that could have been associated to several user profiles.

We observe that the reusability of these data marts is as a result of an inter-dependent relationship among the user profiles. This relationship could be referred to a user group profile. For example, in a departmentalized organization with individuals in each department working towards the departmental objective that has been formulated towards the overall organization goal. Each member of staff in a department has his responsibilities that are inter-dependent with one another. Each member's function or responsibilities will determine his profile and the departmental responsibilities determine the departmental profile within the organization. Furthermore, marketing staff in an enterprise is associated with a marketing profile as a group. But, individuals could belong to corporate, retail and etc marketing responsibility which determines their profile within the department.

At this point, the departmental data mart comes into play: it could be a virtual or a distributed or completely materialized. Several views overlaid on each other enterprise wide could give rise to the enterprise data warehouse. In [22], a data warehouse has been referred to a collection of views overlaid on each other, that is, in good practice we can generate data marts from existing data marts [15]. This could give the watcher in an enterprise the freedom to create views in order to obtain relevant information to aid the process of resolving a particular decision problem. Therefore, there is a need to ensure the accessibility and reusability of the existing data marts.

4 Heterogeneous Information Integration Architecture.

In a proposed architecture for information sources integration as described in [18], three perspectives were identified: a conceptual perspective, a logical perspective and a physical perspective. The enterprise data model is a conceptual perspective representation of the global concept and relations that are of interest to the application [5]. However, it has been observed in [27 and 24] that this model often does not exist in reality and are assumed as problematic. Instead, individual user requirements should be the starting point of the method therefore we have been able to consider this in the developed SITE models of Economic Intelligence processes MEPD [6], MIRABEL [16] and MORPRIE [2].
Serious effort has been made to conceptualize the enterprise model in terms of the decisional problems to be resolved by the economic actors. These decisional problems leads to information search problem. MEPD modeled the decision marker as a person who is capable of identifying the problems to solve in terms of stake, of risk, and of threat which weighs on the enterprise [14]. On the other hand, MIRABEL insists on the clarification of the problem by the person who sighted or announced the problem.
In [18], the integration effort follows the "local as view" approach while our approach is the "global as view" approach. The global as view approach considers the content of the data warehouse as views on the information sources. It can also be used as a description of the data flow from the sources into the integrating system.

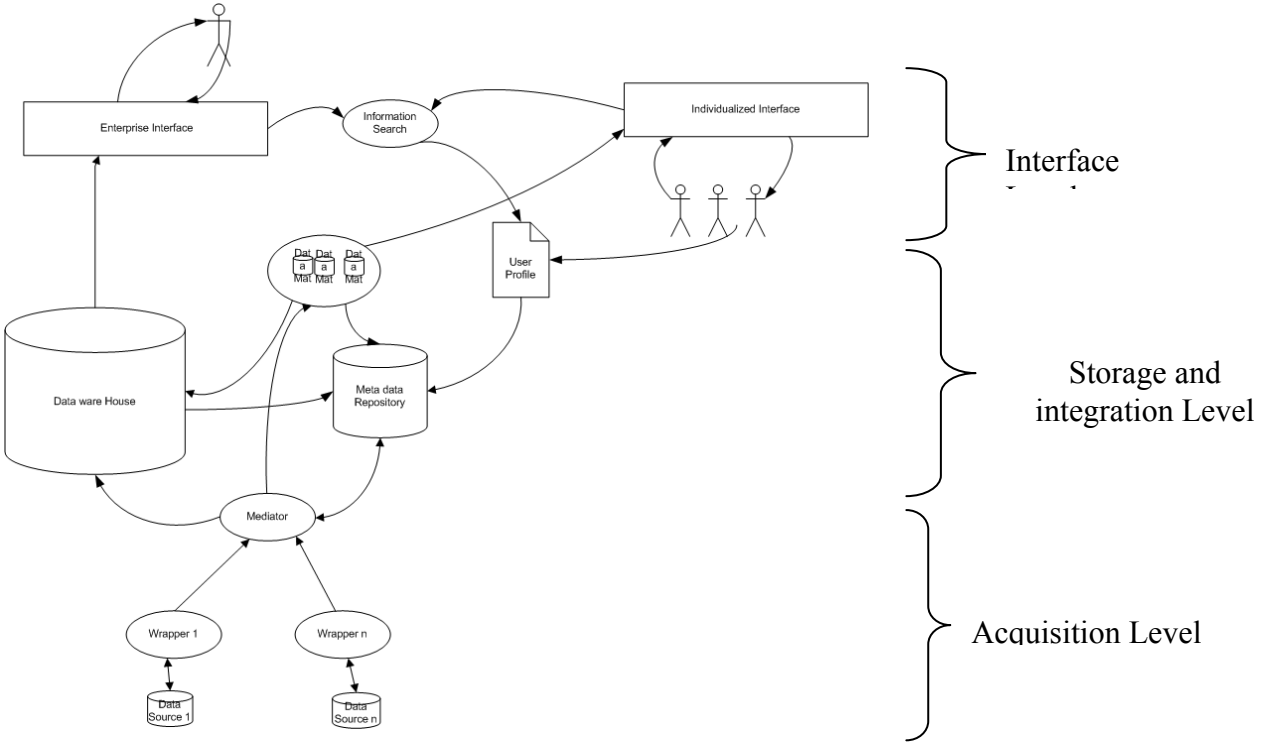

Figure 1. Architecture of materialized information integration.

Keys:

↔ Information flow process

user

We propose architecture for materialized heterogeneous information integration in EI depicted in figure 1. There are three levels identified in the architecture:

- Interface level, constituted by individual and the enterprise interaction level which will provide an adaptive accessibility to the system. This level is concerned for searching information according to user profile. It is based on data marts and meta data repository. Data mart is a data base constructed for a function or a part of enterprise (by example marketing section). Meta data repository contains data about data, by example the conceptual schemas of the different data marts.
- Aggregated information storage and integration level, constituted by the information mediator, data marts, enterprise data marts (Data Warehouse), a metadata repository and users' well expressed information search representations. This contains the metadata and the actual data stores of the system in order for quick exploitation and information reuse. The different sources for interface level are bound to the data warehouse. For us data warehouse is, as Inmon [17] "A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process". Inmon defined the terms in the sentence as follows: **Subject Oriented** (Data that gives information about a particular subject instead of about a company's ongoing operations); **Integrated** (Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole); **Time-variant** (All data in the data warehouse is identified with a particular time period); **Non-volatile** (Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business).
- Endeed the third level is the source acquisition level, which consists of information sources wrappers and the information sources. It is based on two tools : wrapper and mediator which are explained by figure 2.

In EI, we are able to classify information sources as either an open sourced or a formal source. This classification is a based on the method of accessibility and the nature of the sources. We define open source as a source of information without any form of accessibility constraint and redistribution of his content while the formal sources could be databases of an enterprise or/ and the web. In the formal sources serious attempts are made to control its authorship and access. Majority of the online transactional processing systems belong to the later classification with access to their contents could be of diverse levels of restriction. It is a common practice for formal information sources to be protected at specific access levels. This accesses levels are usually modeled and implemented based on the user's function or activity with the enterprise. For example, within a single relational database management system there could be levels of accessibility restriction drilled down to specific tables and the components of such table (Columns and Rows). We have observed some levels of restriction in production databases to the extent of data access at the column levels of a table [23].

Here, we are confronted with the problem of resolving accessibility of disparate information sources whose contents could be involved in resolving a particular decision problem. Also, the problem of selection and mapping of relevant information sources to a user profile. We observed that, to achieve meaningful information integration a special effort should be made to resolve this problem of access and presentation heterogeneity. In our architecture figure1, we propose a context aware wrapper for each information sources.

Figure 2 shows an integration system with n source with n corresponding wrappers. Each wrapper is built with the intention to extract the content of its source to be transformed to an intermediate representation format that can be processed by a mediator. This intermediate representation format for information exchange between the wrappers, and the mediator should be standardized in order to resolve the problems of misnomers, homonyms, and hyponyms and inadequate description of the structure of indentified entities or objects among the retrieved information sets. We

observed that this problem could be resolved at the moment by user's intervention since the interpretation attributed to the concept represented by the identified entities can only be understood by the users in the context of use of the information.
This figure explains the construction of data (or information) base from different sources. Each source is analysed by a wrapper which verifies data quality, non redundance, consistance etc. These different views are merged and transformed into an "immediate format". This format can be, by example, XML which is the current format for exchange, representation and storing data. Then the mediator completes this work to make merge of the different controlled data sources. At the end of the process we obtain a base of integrated information, by example a data warehouse, and, also, meta data repository.
Though, effort is being made by the w3c consortium to represent information in a machine readable and understandable format, "yet this simple idea, however, remains largely unrealized" [25]. The w3c consortium desires to achieve this through the adoption of common conceptualizations refers to as ontologies. The use of ontology is not be discussed in this paper.
Heterogeneity of accessibility methods would have been resolved through the wrappers since it will have unrestricted access (commonly refers to administrative right) to all the content and elements in the source. And problem of transformation between querying languages would also been resolved as each wrapper is built to extract information in the query language of its source. Furthermore, the wrapper would also extract the metadata of the information source and represent it in the same intermediate format employed in the representation of the content before being stored in a metadata repository.
Metadata is defined as the data that describes data.
We decern two types of meta data :

- Structural meta data are meta data which describe the structure of data warehouse;
- Accessibility meta data are specific to data warehouse context. They represent by example the aggregation rules (we wish different orders during time, day, week, month, year...) or where we can find the original data (what source?).

We have added to these types of meta data: user meta data, which describes the knowledge about final user [21].
In order to reduce ambiguity in the content of our information repository, we propose a metadata repository that will store the definition of each element in the information repository according to each user profile. This repository will also, store the definition of the source systems, the definition of user profile and the repository usage (which data marts are used and by who). We are of the opinion that a user of the system should be able to explorer the contents of the metadata repository in order to have an insight to the structures of the source systems, the matching, combination, and merging of their entities which is the result of the integrated information presented to him by the system.

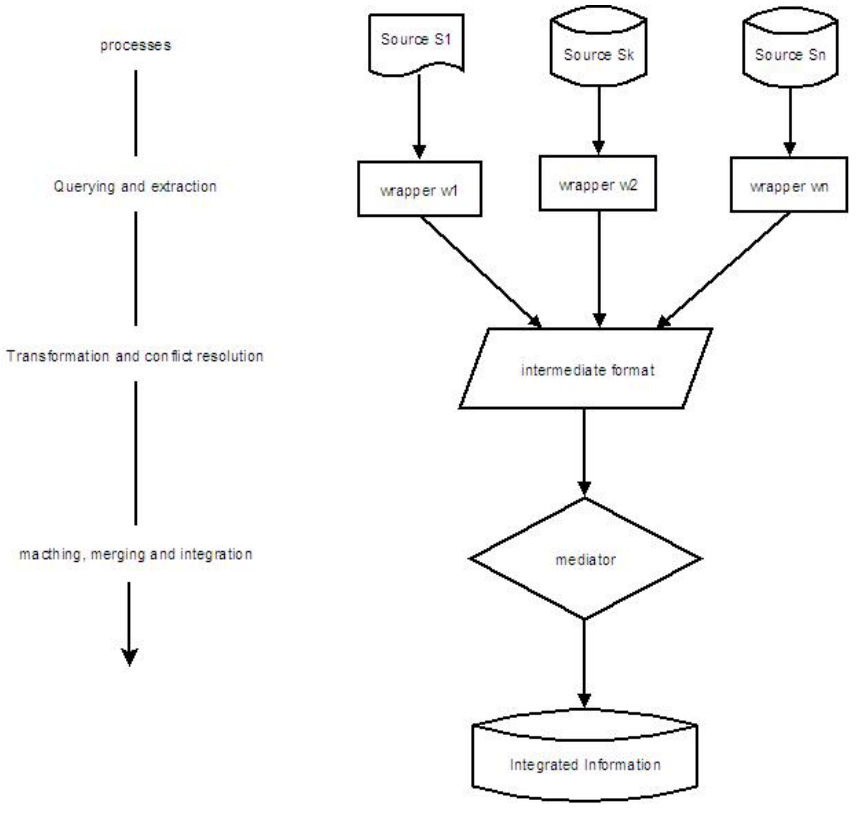

Figure 2: An integration system with n sources and wrappers

5 Conclusion.

In this paper, architecture for materialized heterogeneous information integration was introduced. The proposed architecture puts into consideration the user's information need in its information integration effort. The focus lay on the composition of sources to meet individual user's (or user groups) information requirement, which is used for the incremental build of the enterprise wide data warehouse. This work is still an on going research work, therefore we intend to implement a prototype for the validation of the architecture.

6 Bibliography

[1] **Afolabi B., Goria S.,** Corporate Information Systems Architecture for Business Intelligence Solutions, in Business Intelligence Systems Conference MIPRO 2006, Opatija, Croatia, May, 2006.

[2] **Afolabi B.**, Modeling and adapting the structure of a business intelligence system by observing the behaviour of the user, Ph.D. thesis, Information Sciences and Communication, University Nancy 2, March 2007.

[3] **Afolabi B., Thiery O**., Using Users' Expectations to Adapt Business Intelligence Systems", in G. Budin, C. Swertz et K. Mitgutsch (Eds.), Advances in Knowledge Organization (ISKO), vol. 10, Ergon Verlag, Würzburg, Austria. p. 247 - 254. June 2006.

[4] **Arens Y., Chee C.Y., Hsu C., Knoblock C.A**., Retrieving and integrating data from multiple information sources: Journal of intelligent and cooperative Information systems, 2(2), pp. 127 – 158, 1993.

[5] **Boehnlein M., Vom Ende U.,** A Business Process Oriented Development of Date WareHouse Structures, In Proc. Of Data WareHousing 2000, Physica Verlag, 2000.

[6] **Bouaka N.**, La modélisation du problème décisionnel - modèle MEPD, thèse Université Nancy 2, Novembre 2004.

[7] **Bouaka N., David A., Thiery O.,** Contribution of to the understanding of explanatory factors for a decision-maker problem within the framework of Economic Intelligence, SCI'2002, Orlando, Florida, USA, July 2002.

[8] **Bouaka N., David A**. A proposal of a decision-maker problem to for a better understanding of information needs. 1st International Conference on Information & Communication Technologies: from theory to Applications - ICTTA'04. (Damas, Syrie). 2004, 5p.

[9] **Chawathe S., Garcia-Molina H., Hammer J., Ireland K., Papa-konstantinou Y., Ullman J., Widom J**., The TSIMMIS project: Integration of heterogeneous information sources, In proceedings of IPSJ Confrenec, Tokyo, Japan, pp. 7 – 18. 1994.

[10] **Chu E., Akanksha B., Chen T., AnHai D., Naughton J**., A relational Approach to incremental Extracting and Querying structure in Unstructured Data, In the proceedings of the ACM VLDB'07, pg. 1045 – 1056, September 23-28, 2007, Vienna, Austria.

[11] **Codd E. F., Codd S. B., Salley C. T**., Providing OLAP (on-line analytical processing) to user-analyst: An IT Mandate, Technical report, 1993.

[12] **Collet C., Huhns M.N., Shen W**., Resource integration using large knowledge base in Carnot, IEEE Computer, 24(12), pp. 55 – 62, 1991.

[13] **David A., Thiery O**., Prise en compte du profil de l'utilisateur dans un système d'information stratégique, VSST'01 veille scientifique stratégique et technologique, Barcelone, 2001.

[14] **David A., Thiery O**., Application of "EQUA^{2}te" Architecture in Economic Intelligence, In the proceedings of the conference of Information and Technology Applied to Economic Intelligence, University of Ibadan Nigeria, pp. 32 – 38, August, 2002.

[15] **Goeken M., Knacksted R**., Method for user oriented modeling of data warehouse systems, ICEIS 2006, 2006.

[16] **Goria, S., Geffroy P**., Le modèle MIRABEL : un guide pour aider à questionner les Problématiques de Recherche d'Informations, In Veille Stratégique Scientifique et Technologique - .VSST'2004 (Toulouse). 2004. 5p.

[17] **Inmon W. H**., Building the Data Warehouse, 2nd edition Wiley, 1996.

[18] **Jarke M., Lenzerini M., Vassiliou Y., Vassiliadis P.,** Fundalmentals of Data Warehouses. Second Edition, pp. 41 – 43, September 2002.

[19] **Philippe K.,** La modélisation du problème du veilleur – le modèle WISP, thèse Université Nancy 2 Novembre 2007.

[20] **Michelon M., Knoblock C.A**., Creating relational data from Unstructured and Ungrammatical Data sources, Journal of Artificial Intelligent research, 31(1), pp 543 – 590, 2008.

[21] **Peguiron F., Thiéry O**., Recensement des besoins de l'utilisateur d'un SIS universitaire via un entrepôt de données en open source, Congrès VSST'07 Marrakech Octobre 2007.

[22] **Quix C., Gtaziu S., Jeusfeld M., Staudt M., Vassiliou Y.,** Repository Support for Data Warehouse Evolution: In Proceedings of the International Workshop On Design And Management Of Data Warehouse (DMDW' 99), Heidelberg, Germany, 14. - 15.6. 1999.

[23] **Sang H. Son**, Database Security Issues for Real-Time Electronic Commerce Systems, In IEEE Workshop on Dependable and Real-Time E-Commerce Systems (DARE'98), 1998.

[24] **Schelp J.,** modellierung multidimensionaler datastrukturen analyseorientierter information system. Dr. Universitat Bochum. 2000. (article translated from German to English using – translate.google.com).

[25] **Shadbolt N., Hall W., Berners-Lee T**., The Semantic Web Revisited, Accessed at eprints.ecs.soton.ac.uk/12614/01/Semantic_Web_Revisted.pdf, last accessed 26th October, 2008.

[26] **Shneiderman B.,** Universal usability: Pushing human-computer. Interaction research to empower every citizen, Communications of the ACM 43, 5 May 2000.

[27] **Stahlknecht P., Hasenkamp U**., Einfuhrung in die Wirtschaftsinformatik, 11th edition Berlin et al., Springer, 2005.

[28] **Zhou G., Hull R., King R., Franchitti J.**, Data Integration and warehousing using H20, IEEE data Engineering Bulletin, Special Issue on Materialized Views and data warehousing.