Towards an Agent-Based Approach for Multidimensional Analyses of Semantic Web Data

Salma EL HAJJAMI, Mohammed BERRADA
Laboratoire d'Informatique et Physique Interdisciplinaire (LIPI),
ENS-Fès, USMBA
Fez, Morocco,
salma.elhajjami@usmba.ac.ma
mohammed.berrada@gmail.com

Mostafa HARTI
Laboratoire Informatique, Modélisation et Systèmes (LIMS),
Faculté des Sciences Dhar El Mehraz, USMBA
Fez, Morocco,
mostafa.harti@usmba.ac.ma

Gayo Diallo
Univ. Bordeaux, INSERM,
BPH Research Center,
team ERIAS, UMR 1219, F-33000
Bordeaux, France,
Gayo.Diallo@u-bordeaux.fr

***Abstract*——OLAP analytical systems are essential technologies in decision-making processes; they provide an efficient way to carry out complex analysis in a simpler and faster way to decision-makers. In today's dynamic and competitive business contexts, the stored internal data within companies does no longer provide enough information for decision-making processes. Therefore, decision analysis systems could be improved by including external data available through the semantic web in order to provide multiple perspectives to decision makers. In this article, we describe a preliminary approach based on the use of multi-agent systems for multidimensional analysis of external data coming from the semantic web also gives a short review of recent research works combining business intelligence and semantic web technologies. The proposed approach is based on an evolutionary architecture by dint of the "agents" technology. The different stages of the analysis are considered tasks that will be assimilated to services, managed by agents.**

Keywords— OLAP; semantic web; business intelligence; multi-agent system; multidimensional analysis.

I. Introduction

Business Intelligence (BI) is intended to collect, transform and synthesize data available from existing sources to generate appropriate analysis information for decision-making tasks. Companies (in the broad sense of the term, private companies, public institutions, organizations, etc.) have realized the efficiency of On-line Analytical Processing (OLAP) in the analysis and exploration of data. This technology has been the subject of extensive scientific research and has proven itself in the areas of data management and knowledge extraction from static, well-structured data. The most appropriate way to facilitate on-line analytical processing is the multidimensional data modeling that allows data to be represented as points in a multidimensional space. Multidimensional modeling is therefore a technique that aims to organize data in such a way that OLAP applications are efficient and effective. However, this technique is not adapted to what is called semantic data. The advent of the Semantic Web (SW) has largely contributed to the emergence of these kind of data [2], which leads to the so called linked data. The SW was designed as a way to build semantic spaces on the published web content so that information can be effectively retrieved and processed by humans and machines for a wide variety of tasks. Moreover, it is intended to become a major source of information for business intelligence. However, this source of information is currently under-exploited, because the exploration of linked data raises many problems, notably in terms of structuring and storage on the one hand and analysis on the other hand.

Moreover, in today's highly competitive business environment, additional information from outside an organization, mainly on the Web, should also be included in analyzes in order to provide multiple perspectives for decision makers [1]. A significant amount of information can be found in external data sources such as web portals, social medias, data stores, whether unstructured or less structured, as product reviews, customer complaints, and so on. Since it is now clear that the importance of using this kind of additional data is established for the decision-making process, the need to manage and process them has steadily increased. Companies have started to look into these rich sources of information in order to increase their profits and improve their products and services [1]. Thus, the idea is to explore all these new types of data and include them in their OLAP analysis. However, these data are dispersed and managed under different schemas. During an analysis process, decision makers should navigate among the different sources to gather all the useful information. The dispersion of SW data leads to repetitive look-up to obtain relevant information. This reduces the efficiency of the analysis.

The objective of this article then is first to present an overview of research works conducted on Business Intelligence domain coupled with Semantics and propose an approach that combines the domain of BI with SW technology.

The remainder of this paper is structured as follows. Section 2, deals with the state of the art of related works about BI and SW. we study agents based semantic data management in the section 3. Before to conclude and outline some perspectives to this research work, we introduce in section 4, our approach for the multidimensional analysis of SW data.

II. RELATED WORK

A. Semantic web technology overview

The semantic Web expression attributed to Tim Berners-Lee [2] within the W3C, refers first to the vision of an optimized web representing the largest space for the exchange of resources between humans and machines allowing a qualitatively superior exploitation of large volumes of information and services available on the web. The purpose of the semantic web would therefore be to make explicit the semantic content of resources in the web (documents, web pages, services, etc.) by the machines and software agents so that they can understand the meaning of the information via the descriptions Contained in these resources. The current vision of the semantic web proposed by Berner Lee is represented in a pyramidal architecture in several different layers Fig.1

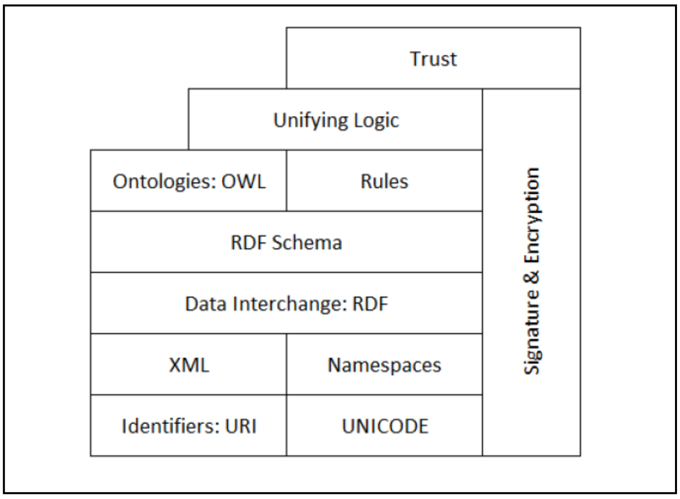

Fig. 1. Semantic web pyramid architecture [W3C]

The XML extensible markup language provides the correct syntax for modeling the content of data. The Resource Description Framework (RDF) [3] is a graph-based data model that describes objects (resources) and relationships between them. The RDF Schema [4] describes the properties and classes of RDF resources. OWL [5] adds additional vocabulary support to describe properties and classes, relationships between classes, cardinality, equality, richer property typing, property characteristics, and hierarchies of properties and classes. Traditional web technologies focus mainly on the representation of data, while the Semantic Web allows the possibility of enriching the traditional Web using "semantic" information, understood by machines that will facilitate research and the use of Web resources (Web pages, images, services, etc.).

B. Semantic web data analysis

The rapid evolution of new technologies, especially the Web, has given rise to new problems related to SW data, in particular regarding their structuring and storage on the one hand and their analysis on the other. These data pose several challenges for BI, especially since they are very rich in information but complex, which makes their processing difficult for decision-making. In this context, a particularly domain of interest is the on-line analytical processing of linked data on the SW, are likely to contain a significant amount of information useful for decision makers.

Multidimensional analysis of linked data has been the subject of extensive work, most of which concerns the storage of data extracted from the SW in a local data warehouse. In this way, OLAP analysis can be performed on these extracted data using existing analysis tools. However, this process is performed mostly in a semi-automatic, they are not able therefore to efficiently handle the growing and dynamic nature of the SW in order to perform real-time analysis [6]. Alternative approaches have dealt with multidimensional analysis on linked data on the SW without an Extraction, Transformation and Loading (ETL) process, by proposing Frameworks based on the RDF Data Cube (QB) vocabulary. the latter has been specially designed to publish cubes of data on the Web [7][8][9] [10].

The first axis focuses on extracting, transforming, and loading data from the semantic web into a local data warehouse so that decision makers can directly perform their OLAP analyses[19][20]. Its main advantage is the ability to reuse existing OLAP tools while analyzing linked data transformed into an OLAP cube. However, it presents some limitations in particular the storage of extracted data into a local data warehouse. In addition, since these approaches are semi-automatic to perform real-time analysis as they cannot efficiently handle the dynamic and changing nature of information published on the Web. Therefore, the second research axis attempts to overcome the disadvantages of this one, by providing an OLAP analysis without ETL, this is possible by, directly analysing the data of the SW thanks to frameworks based on the RDF Data Cube (QB)[1] vocabulary (Fig.2).

[1] http ://www.w3.org/TR/vocab-data-cube.

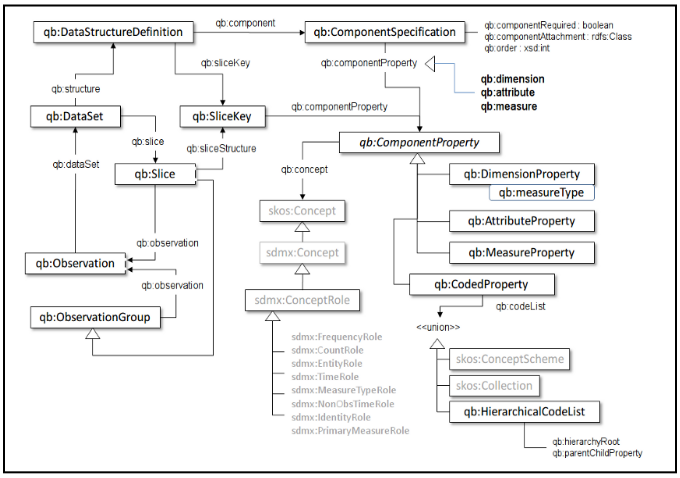

Fig. 2. The QB Vocabulary

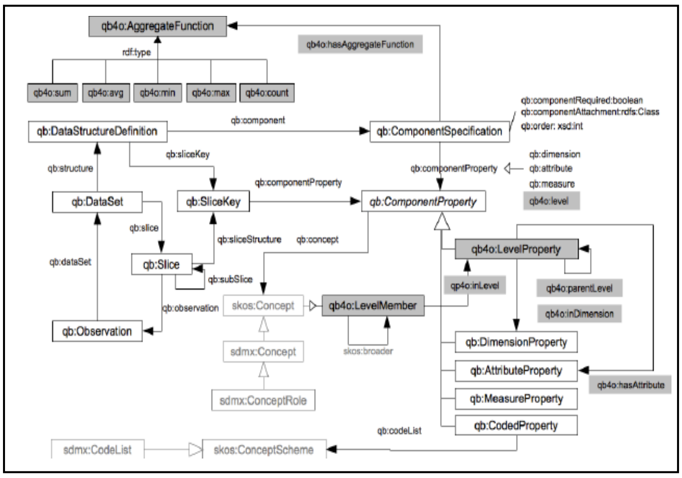

Fig. 3. The QB4OLAP Vocabulary

C. Frameworks for multidimensional modelling and analysis

Multidimensional modeling aims to organize data in such a way that OLAP applications are efficient and effective. Most current Frameworks are based on the RDF Data Cube (QB) vocabulary, which was specifically designed to publish data cubes on the Web. QB does not provide a mechanism to represent several levels on one dimension and the relationships between the levels of the schema. In response to that, Kämpgen et al [7] defines QB-like extension of the model QB in order to represent statistical data in a multidimensional model. They describe how to perform OLAP analyzes on data published in QB using the SPARQL[2] query language which is a W3C standard for querying RDF data. However, their solution does not support the multi-level hierarchical structure, and multiple hierarchies in a dimension. Etcheverry and Vaisman [8] introduces a new RDF vocabulary called Open Cube (OC) for multidimensional modeling of RDF data. OC provides a set of classes and properties to model the different structures of the multidimensional model (dimensions, attributes, measures), including hierarchical relationships between dimension attributes. From the RDF collections described using OC, different OLAP operations can be performed directly via queries expressed in SPARQL. Although this solution is based on a multidimensional modeling of the RDF data and allows expressing the OLAP operations in terms of SPARQL, its main limit is the re-use of the data already published in QB (which is standardized). The same authors [9] introduces the QB4OLAP vocabulary (Fig.3). The latter extends and remains compatible with QB in order to support the multidimensional modeling of linked data. In [10], an extension of QB4OLAP that supports several hierarchies in a dimension and cardinalities between level members is proposed. In addition, mechanisms to transform an existing relational data warehouse to the QB4OLAP is provided.

We have summarized in Table I, vocabularies that allow the modeling of the multidimensional data of the SW in order to perform an OLAP analysis. The vocabulary QB-like [7] does not have sufficient capacity to manage OLAP. (OC) is a modeling language that supports multiple hierarchies in a dimension, however the data already published in QB (which is standardized) cannot be reused by OC. The vocabulary of QB4OLAP is an extended version of QB, offering more features to support OLAP.

TABLE I. COMPARAISON OF VOCABULARIES

Multidimensional Data Concept	QB-like	OC	QB4OLAP
Dimensions	✓	✓	✓
Levels	✓	✓	✓
Level members	✓	✓	✓
Rel. between level members and levels	✓	✓	✓
Roll-up relations between Levels	✓	✓	✓
Roll-up relations between Level Members	✓	✓	✓
Multiple hierarchies in a dimension		✓	✓
Measures	✓	✓	✓
Multiple measures in a cube	✓	✓	✓
Aggregation functions	✓	✓	✓
Ability to reuse data in QB	✓		✓

[2] //www.w3.org/TR/sparql11-query/

III. AGENTS BASED SEMANTIC DATA MANAGEMENT

The management of data on the web consists in a series of tasks to be performed, so it is necessary to use a cooperative system so that these different tasks are carried out in a coherent way. From this point of view, it is natural to introduce the notion of agent. A software agent is a conventional program called intelligent. Multi-Agent System (MAS) refers to a more or less extensive set of actors that communicate with one another [11]. This set aims at the realization of a well-defined task, where each agent has a specific objective and offers services to perform some tasks, autonomously and communicates the results obtained to a receiving actor (human or software). Much work has been done to manage data on the Web using multi-agent systems; The InfoSleuth [12] project extends the capabilities of the Carnot technologies developed at MCC in dynamically changing environments, which is specialized in integrating heterogeneous data sources. InfoSleuth is accomplished by collaborative agents, and the use of Java as a common agent wrapper, with the aim of automating the collection and analysis of dynamic data distributed over the Web, where each agent is an autonomous process that specializes in a service. A group of agents collaborate with each other to accomplish complex tasks of collecting and analyzing dynamic data. Ontologies are also used to impose a uniform view on data from different sources. Authors in [13] proposed an innovative approach for the management of an organizational memory combining ontology engineering, SW technology, and multi-agent systems in an integrated solution. The keystone of [13] is a common and shared ontology [14] ensuring the coherence of memory and communications between agents. Authors in [15] present a SW services discovery architecture using agent technology and ontologies. This architecture integrates software components and exploits a domain ontology that is used in the discovery phase of Web services. It facilitates the automatic discovery of services since it allows to refine the search process that matches a request and offers of services. In [16] an approach for SW service using domain ontological retro-engineering to generate semantic links is described. This approach consists of two phases: the extraction of relevant information, and the analysis of the extracted information based on a domain ontology using a similarity criterion. [16] allows a classification of the WSMO SWS per the domain of ontology. In [17], they proposed an agent-based architecture for the management of various data sources in a dynamic environment (Fig.4). This work, based on [12], follows a mediator based approach for managing RDF-based sources. The mediator approach allows a virtual integration, meaning that no central data store is built. In addition, it integrates linked data sources while maintaining their local autonomy and allows SPARQL based queries. Our approach is inspired by this latter work and we have adapted an agent-based architecture for the selection and retrieval of linked data on the SW.

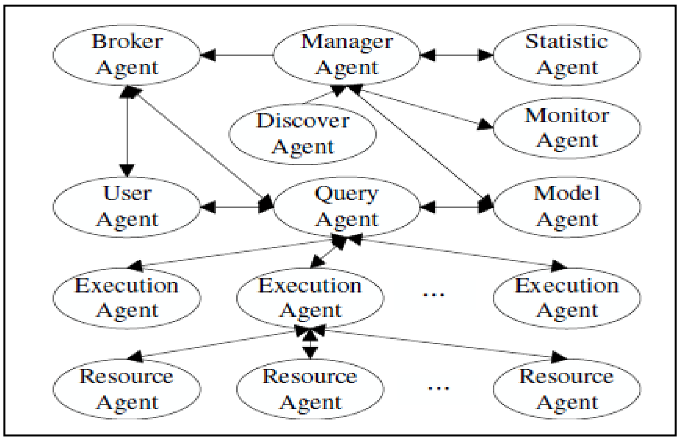

Fig. 4. LDMS Architecture

IV. PROPOSED APPROACH

By comparing the two main categories of approaches for OLAP analysis of linked data, we observe that approaches based on direct manipulation of RDF data are more advantageous in terms of flexibility and adaptation to the specificities of data published on the Web. However, its main disadvantage is related to the materialization of data. Indeed, large datasets from the SW are queried on the fly. These data without a cleaning process can bring inaccurate information to decision makers [18].

A. Detail of our approach

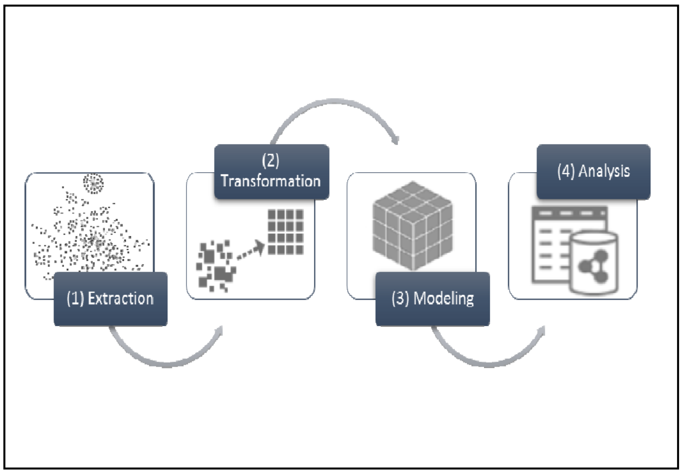

Fig. 5. The steps of the proposed approach

Our approach consists in proposing a methodological framework and tools for the extraction, transformation, multidimensional modeling of the data and finally their exploration by OLAP techniques. The core of the approach revolves around a process based on multiagent systems, where the different tasks in question will be assimilated to the services offered by different agents. The different SW components are presented in Fig.5.They are detailed as follows:

1) *Extraction of the linked data:* The choice of the data to be observed is a real handicap for the analysis of linked data on the SW. OLAP analysis of various sources and types (RDF, RDFS, OWL ...)

2) *Transformation of extracted data:* This task is performed by the agents in order to prepare extracted data for analysis and to represent them in a unified data format

3) *Modeling of the multidimensional data schema:* This task is managed by agents whose role consists in modeling a multidimensional schema in order to offer a framework for analyzing data by observing facts through axes of the dimensions and axes of the measures. To do this we opt for the QB4OLAP vocabulary because of its expressiveness and it is ability to perform OLAP analyses. Thus, the construction of the cubes of data is done on the fly according to the end user analysis needs.

4) *OLAP multidimensional analysis*: the complexity of the linked data requires specific aggregation operations. This task is performed by the agents in order to translate OLAP operators to SPARQL queries and implementing them on the defined cube.

B. Architecture of our approach

Our proposed approach is based on a scalable architecture offering a great flexibility and a strong structuring based on a platform of agents. The architecture depicted in Fig.6 below consists mainly of three components:

- User interaction interface: through which the user interacts with the system.
- Intelligent agents: which communicate and interact dynamically in order to achieve a complex goal.
- Data sources: which are external data sources from the SW, which is a very rich information resource.

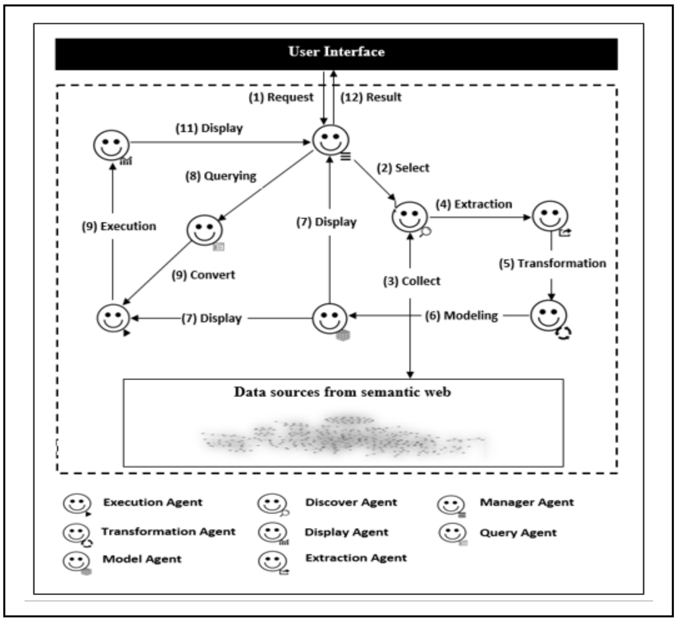

Fig. 6. Architecture of the proposed approach

We have instantiated eight agents offering services allowing the multidimensional analysis of the linked data on the SW in which each of these services generates products:

- Manager Agent: This is the system driver. It acts as an interface between the system and the user. It is responsible for overseeing agent migrations.
- Selection Agent: Its role is to collect the data according to the need of the analysis.
- Extraction Agent: It allows extraction from the collected data.
- Transformation Agent: Its goal is to transform the extracted data into a unified format.
- Model Agent: It allows to model a multidimensional schema
- Query Agent: Its role is to translate the OLAP operators chosen by the user into SPARQL queries.
- Execution Agent: Used to execute SPARQL queries.
- Display Agent: Its role is to display the result of the SPARQL queries.

C. Sample Scenario

In the following, a sample scenario is used to demonstrate how agents in the presented architecture interact with each other. During system start-up, A user submits a query through the user interface provided by *ManagerAgent* (1). After verification, the request is sent to *SelectionAgent* (2), which collects the data according to the need of the analysis (3) and transmits them sequentially to *ExtractionAgent* (4), the latter makes the extraction and the cleaning of the collected data and transmits it to *TransformationAgent*, which transforms the extracted data into a unified format (5) and forwards it to *ModelAgent* (6), which takes care of multidimensional modeling and generates a cube. *ManagerAgent*, *ExecutionAgent* retrieve the created multidimensional cube (7). Then *QueryAgent* translates the OLAP operators chosen by the user into SPARQL queries (8), *ExecutionAgent* retrieves the SPARQL queries (9) and executes them on the multidimensional cube created (10) and passes the result to *DisplayAgent*, which is responsible for presenting the Query results and passes them to *ManagerAgent* (11). To finish, *ManagerAgent* displays the results (12) according to the user-selected queries (7) (11).

V. CONCLUSION AND PERSPECTIVES

The SW is considered to be a major source of information for business intelligence. In order to take advantage of this new source, the domain of BI must be enriched by new approaches and methods to process the data coming from external sources, mainly from the SW, and also include them in the analyzes to provide multiple perspectives to decision makers. We have outlined an approach which combine BI techniques and SW technologies to carry out a multidimensional analysis of SW data. It consists in a methodological framework and tools using multi-agent systems for extraction, transformation, multidimensional modeling OLAP on-line analysis techniques. This approach is based on an evolutionary architecture thanks to agents technology, where these different stages are considered as tasks assimilated to services, managed by agents.

For the perspectives of this preliminary work, as our approach is in the beginning step, we plan to develop and implement the prototype of the proposed framework and testing the solution with a vital domain such as the Social Web.

REFERENCES

[1] A. Abello, O. Romero, T. B. Pedersen, R. Berlanga, V. Nebot, M. J. Aramburu, and A. Simitsis, “Using semantic web technologies for exploratory OLAP : a survey”, IEEE Transactions on Knowledge and Data Engineering, 27(2), 571-588, 2015.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, “The semantic web”. Scientific american, 284(5), 28-37, 2001.

[3] K. Graham, C. Jeremy, “Resource Description Framework (RDF) : Concepts and Abstract Syntax”, 2004

[4] D. Brickley, and R. V. Guha, “RDF vocabulary description language 1.0 : RDF schema”, http ://www.w3.org/tr/2002/wd-rdf-schema-20020430/, 2002.

[5] M. Dean, G. Schreiber, “OWL Web Ontology Language Reference”, http ://www.w3.org/tr/owl-ref/, 2004.

[6] O. Romero, and A. Abelló, “Automating multidimensional design from ontologies”, In Proceedings of the ACM tenth international workshop on Data warehousing and OLAP, pp. 1-8, ACM, November 2007.

[7] B. Kämpgen, S. O’Riain, and A. Harth, “Interacting with statistical linked data via OLAP operations”, In Extended Semantic Web Conference, pp. 87-101, May 2012, Springer Berlin Heidelberg.

[8] L. Etcheverry, and A. A. Vaisman, “Enhancing OLAP analysis with web cubes”, In Extended Semantic Web Conference, pp. 469-483, May 2012, Springer Berlin Heidelberg.

[9] L. Etcheverry, and A. A. Vaisman, “QB4OLAP : a new vocabulary for OLAP cubes on the semantic web”, In Proceedings of the Third International Conference on Consuming Linked Data, Volume 905, pp. 27-38, November 2012. CEUR-WS. org.

[10] L. Etcheverry, A. Vaisman, and E. Zimányi, “Modeling and querying data warehouses on the semantic web using QB4OLAP”, In International Conference on Data Warehousing and Knowledge Discovery, pp. 45-56, September 2014, Springer International Publishing, 2015.

[11] K. Sykara and D. Zeng and K. Decker and A. Pannu and M. Williamson,” Distributed Intelligent Agents”, IEEE Expert, pp 1-32, 1997.

[12] R. J. Bayardo Jr, W. Bohrer, R. Brice, A. Cichocki, J. Fowler, A. Helal, ... and M. Rashid, “InfoSleuth : agent-based semantic integration of information in open and dynamic environments”, In ACM SIGMOD Record, Vol. 26, No. 2, pp. 195-206, June 1997, ACM.

[13] F. Gandon, R. Dieng-Kuntz, O. Corby, and A. Giboin, “Web Sémantique et Approche Multi-Agents pour la Gestion d’une Mémoire Organisationnelle Distribuée”, In 13eme IC, Conférence Ingénierie des Connaissances, May 2002.

[14] D. Fensel, “Ontologies”, In Ontologies (pp. 11-18), Springer Berlin Heidelberg, 2001.

[15] B. Chafik, and K. Okba, K, “Une Approche basée agent pour la Découverte de Services Web”, In CIIA, Eds. CEUR Workshop Proceedings, vol. 547. CEUR-WS.org, 2009.

[16] B. Fatima, H. Abdelkader, and B. Djelloul, “Description et Classification des Services Web Sémantiques”, Nature and Technology Journal, Issue 10, Volume A January, 2014.

[17] X. Li, Z. Niu, and C. Shi, “An agent-based linked data integration system”,TheFifthInternationalConferenceontheApplicationsofDigital Information and Web Technologies (ICADIWT 2014), Bangalore, 2014, pp. 113-117.

[18] S. Laborie, F. Ravat, J. Song, and O. Teste, “Combining business intelligence with semantic web : Overview and challenges”, Actes du XXXIIIème Congrès INFORSID, Biarritz, France, 99-114, 2015.

[19] O. Romero and A. Abelló. “Automating multidimensional design fromontologies”,(2007), In *Proceedings of the ACM tenth international workshop on Data warehousing and OLAP* (pp. 1-8). ACM.

[20] V. Nebot, R. Berlanga, J.M. Pérez, M. J. Aramburu, and T. B. Pedersen, (2009), “Multidimensional integrated ontologies: A framework for designing semantic data warehouses”. In *Journal on Data Semantics XIII* (pp. 1-36). Springer Berlin Heidelberg.