Combining Business Intelligence with Semantic Web: Overview and Challenges

Sébastien Laborie[1], Franck Ravat[2], Jiefu Song[2], Olivier Teste[3]

1. LIUPPA – T2I, Université de Pau et des Pays de l'Adour
2 Allée du Parc Montaury 64600 Anglet
sebastien.laborie@iutbayonne.univ-pau.fr

2. IRIT - Université Toulouse I Capitole
2 Rue du Doyen Gabriel Marty F-31042 Toulouse Cedex 09
{ravat|song}@irit.fr

3. IRIT - Université Toulouse II Jean Jaurès
1 Place Georges Brassens F-31703 Blagnac Cedex
teste@irit.fr

ABSTRACT. *Under today's highly complex and dynamic business environment, external data (most often issued from web) need to be included in traditional On-Line Analytical Processing (OLAP) analysis so that decision-makers would be well-informed before making effective decision. Including external web data requires knowing the exact semantic meaning in order to use the right information at the right time. Semantic Web (SW) technologies allow semantically annotating data so that we can exchange several descriptions over web data, do reasoning over these descriptions and ensure interoperability between humans and systems. A combination of BI technologies with SW will both enhance BI analysis with web data and allow analyzing SW data through BI tools. In this paper, we firstly introduce basic concepts of the BI and SW domains. Then, we present recent research results using SW to enhance OLAP analysis. At last, we identify challenges requiring future research efforts to achieve a complete incorporation of BI with SW.*

KEYWORDS: *Semantic Web, Data Warehouse, Multidimensional Analysis*

1. Introduction

The domain of Business Intelligence (BI) aims to provide a set of tools, methods and technologies for supporting and facilitating decision making. In the context of BI, a data warehouse is used to collect, organize and store subject-oriented, integrated, time variant and non-volatile data (Inmon, 1996 ; Kimball, 1996). Data from different sources (generally internal databases) are periodically added into data warehouse after being cleaned and transformed into a specific structure with the help of Extract-Transform-Load (ETL) process. Traditional BI tools, such as On-Line Analytical Processing (OLAP), have been successfully applied to large amount of data coming from internal databases. However, the dynamic nature of today's business activities forces traditional BI to open its gate to external data in order to answer to more heterogeneous and open analysis scenario (Chen, Chiang, et Storey, 2012). As an increasing quantity of semantically annotated data is available over Internet[1], including Semantic Web (SW) information in traditional OLAP analysis process is a promising way to enhance traditional BI analyses (Trujillo et Maté, 2012 ; Zorrilla *et al.*, 2012 ; Abelló *et al.*, 2013). For instance, a decision-maker may want a better overview of a product by populating a business report with web-published customers' opinions and markets' information (Berlanga *et al.*, 2014).

Even though BI and SW have been two different research directions over the last decades, recent research results show that the convergence of these two domains is inevitable and beneficial for both sides. BI offers powerful tools for analyzing large amount of web data, while SW data have an important density of valuable information that can be used for enriching business analysis (Thi et Nguyen, 2008 ; Kämpgen et Harth, 2011 ; Zorrilla *et al.*, 2012 ; Etcheverry et R. A. Vaisman, 2012 ; Abelló *et al.*, 2013 ; Ibragimov *et al.*, 2014 ; Aufaure et Chiky, 2014).

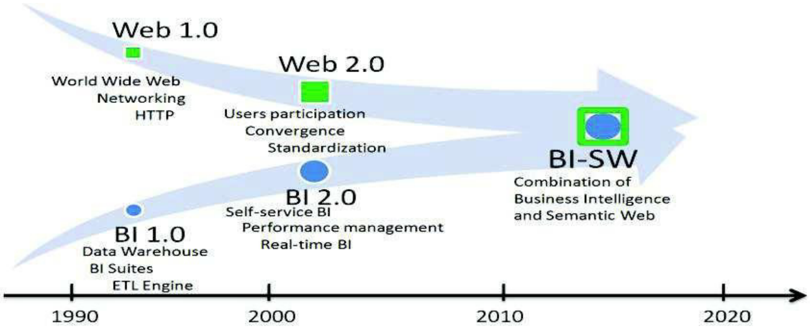

Figure 1. Evolutions in domains of BI and Web

Combining BI with SW, however, is not a trivial task due to the scalability, complexity and heterogeneity of SW data. It raises the following questions: How to integrate heterogeneous SW data in a BI system originally designed for factual data? How to carry out multidimensional analyses over large amount of SW data in the lack of relevant model? How to present analysis results containing both factual data and SW data? These questions are examples of issues waited to be resolved.

The aim of this paper is to present an up-to-date survey of research results and outline future research challenges in BI and SW domains. The rest of the paper is organized as follows. We (i) briefly present the concepts of BI and SW in the section 2; (ii) give an overview of recent research results combining the domain of BI with SW in the sections 3 and 4; (iii) discuss emerging trends and perspectives of future researches in the section 5.

[1] http://linkeddata.org

2. Concepts of Business Intelligence and Semantic Web

2.1 Business Intelligence

The term of *Business Intelligence* (BI) refers to a set of techniques used for collecting, extracting and analyzing business data to support decision-making process. Coming from heterogeneous and distributed operational sources, data used in decision-making process are stored in *Data Warehouse* after going through a process called *ETL* (standing for Extraction, Transformation and Loading).

Among different types of data warehouse, *On-Line Analytical Processing* (OLAP) data warehouse has been a specific research topic for over a decade. The concepts of OLAP were firstly proposed in (Codd, Codd, et Salley, 1993), they provide solutions for creating, managing, analyzing and reporting large amount of multidimensional data in an interactive way. Among all data models proposed for OLAP, the *Star Schema* (Kimball, 1996) is the most widely accepted model (Chaudhuri, Dayal, et Narasayya, 2011). At conceptual level, *Star Schema* presents data according to subjects of analysis (facts) and axes of analysis (dimensions). At logical level, *Star Schema* can be built on top of different types of databases: *Multidimensional OLAP* (MOLAP), *Relational OLAP* (ROLAP) and *Hybrid OLAP* (HOLAP). At physical level, *Star Schema* can be implemented in different ways, as long as the implementation conforms to the twelve evaluation rules defined in (Codd, Codd, et Salley, 1993), such as multidimensionality, transparency, accessibility, etc. Together with the multidimensional data model, a set of operators is indispensable for OLAP analysis. They permit to aggregate information (Drilldown, Rollup), filter analysis results (Slice, Dice) and change analysis axes (Pivot).

(Kimball, 1998) points out that the main advantages of OLAP model lie in its simplicity and understandability that permit users to interact with large amount of complex data in an efficient way. Nowadays, OLAP is a well-mastered technology when it comes to homogenous and structured data in classical data warehouse. However, as factual data provide only limited and partial views over open-world business scenarios (Zorrilla *et al.*, 2012), the data warehouse community looks for solutions for enriching data collection with external data.

2.2 Semantic Web

To accurately exploit web data, a system needs to be capable to read the exact semantic meaning of web-published information. An acknowledged way to publish machine-readable information is to use *Semantic web* (SW) technologies. The purpose of SW technologies is to fix a common vocabulary and a set of interpretation constraints (inferring rules) so as to semantically express metadata over web information and allow doing some reasoning on it. These technologies provide the capability of annotating web data with semantics, e.g., through RDF[2]

[2] http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

and ontologies, hence generating a web of semantic linked data (e.g., Linked Open Data cloud[3]).

Tim Berners-Lee pointed out four principles that SW data should follow[4]: use *Uniform Resource Identifiers* (URIs) to identify object; use *Hypertext Transfer Protocol* (HTTP) to facilitate searching for objects by human-beings; use the *Resource Description Framework* (RDF)[5] format as standard to provide descriptive information about an object; link URIs to others in order to connect individual data into a data web. Compared to traditional web technologies which focus mainly on data representation, SW puts a higher value on providing machine-readable information about web resources and relationships between resources.

More specifically, SW presents human knowledge through structured collections of information and sets of inference rules (Berners-Lee, Hendler, et Lassila, 2001). The basic data model is RDF permitting to express simple statements about resources, using named properties and values (cf. figure 2). Resources described by RDF are not necessarily retrievable on the web, they can be anything with an unique identity, from physical objects to abstract concepts (McBride, 2004). A *Triple Store* permits to store RDF data. The set of statements in a RDF Triple Store is composed of URIs, blank nodes and literals. A RDF triple refers to *subject*, *predicate* and *object*: a subject is a web resource identified by a URI or a blank node; an object can be a web resource or a literal that possesses a primitive value; a predicate is a binary relationship connecting a subject with an object. For instance, in the figure 2 we can find the predicate denoted by the label *Concerns* associating the resource *Sales* with another resource *ProductX*, and another predicate named *hasPrice* connecting the subject denoted *ProductX* to a textual literal "*30"* which is the product's price.

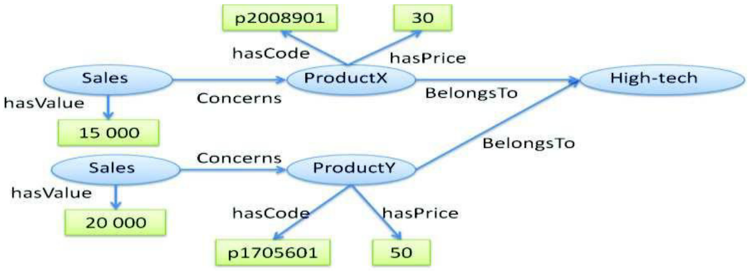

Figure 2. Example of data modeled in RDF format

There exist other SW formats with more powerful expressivity than RDF. Built on top of RDF, RDF Vocabulary Description Language (or RDF schema or RDFS[6]) is a language that defines the terms used in RDF graph. Equivalent to schema

[3] http://lod-cloud.net

[4] http://www.w3.org/DesignIssues/LinkedData.html

[5] http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

[6] http://www.w3.org/TR/rdf-schema/

definition language in relational and object-oriented data model, RDFS is used to describe classes of resources. In other words, RDFS is a simple ontology definition language which allows expressing taxonomies. The concepts of RDFS are described in form of a set of predefined RDF resources with special meanings. However, the reasoning capacity of RDFS is very limited, only basic inferences about taxonomies are supported (Horrocks, Patel-Schneider, et van Harmelen, 2003). Facing to this issue, the Web Ontology Working Group of W3C develops more powerful ontology languages, such as OWL-Lite, OWL-DL, OWL-Full, which allows defining explicit, formal conceptualizations of domain models. In general, OWL enhances the expressivity of RDF and RDFS schema by adding Description Logic (DL). Hence, OWL is an ontology language with sufficient expressive power which can support efficient reasoning through well-defined syntax and semantics (Antoniou et van Harmelen, 2004).

By using the SW formats, web resources can be enriched with annotations and other markups capturing the semantic metadata of resources. However, not all current technologies are fully compatible with the semantic enrichment. For instance, traditional *Information Retrieval* (IR) technologies cannot directly exploit the annotated semantic meaning of web resources (Finin *et al.*, 2005). On the other hand, new research directions have been proposed to combine traditional research approaches with SW technologies, such as *Semantic Information Retrieval* (Fernández *et al.*, 2011), *Exploratory OLAP* (Abelló *et al.*, 2015) etc. In this paper, we only focus on the emerging research direction which aims at enhancing traditional BI with new SW technologies.

3. Overview of researches combining BI with SW

Nowadays, a large number of researches try to merge OLAP analysis with SW technologies both in data integration and data processing levels. This research direction permits to combine powerful tools and technologies in both domains. But it is not a trivial work mainly due to the reason that follows: OLAP requires a specialized data model to support multidimensional analysis over aggregated values of measurements at different granularity levels. However, SW does not dispose of appropriate model fully satisfying criteria about hierarchical levels proposed by (Codd, Codd, et Salley, 1993). Carrying out OLAP analysis directly over SW data is difficult and inefficient by the lack of suitable data model bridging the gap between SW and OLAP domains. Actually, OLAP is originally conceived for analysis over homogenous and stable warehoused data. With arrival of profusion of schema-less Web information, data become more and more heterogeneous and volatile. By mentioning the volatility of SW data we refer to the quick, unceasing and unpredictable changes in SW data sources. Traditional OLAP technologies are challenged while being applied to analyses over SW data.

Facing to these issues, lots of research efforts have been made to combining OLAP with SW. Two types of approaches can be identified (Figure 3). The first approach is OLAP-analyses oriented, which consists of extracting, transforming and then storing multidimensional SW information in traditional OLAP data warehouses

(§3.1), so that it can be analyzed through existing OLAP tools. The second approach is multidimensional modeling oriented, whose aim is to carry out OLAP analyses directly over RDF-like data modeled in an appropriate multidimensional format (§3.2). At the end of the section, we provide a conclusive table (cf. Table 1) that summarizes all mentioned work.

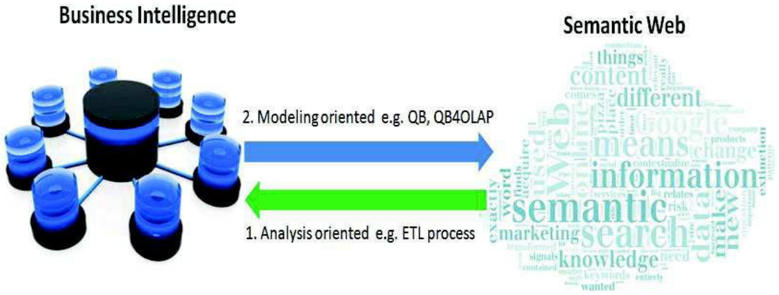

Figure 3.Main approaches to combining BI with SW

3.1 OLAP-analysis oriented approach

OLAP analyses are carried out through analysis operators, such as *roll-up*, *drill-down*, *rotate* and so on (Ravat *et al.*, 2008). Analysis results are usually presented in *Multidimensional Table* (MT) allowing visualizing several analysis axes around a subject. Based on a MT, decision-makers can further carry out OLAP operators to continue their analyses.

OLAP operators are only applicable to specialized data structures (Harinarayan, Rajaraman, et Ullman, 1996 ; Ravat *et al.*, 2008 ; Etcheverry et R. A. Vaisman, 2012), RDF descriptions, however, do not dispose component that can directly support OLAP analysis. For instance, in order to carry out drilldown and rollup operations, we need to represent data according to hierarchical levels within a dimension. However, even though RDF triple can be used to describe web resources and relationships between them (instance level), it does not allow revealing hierarchical relationships within a dimension structure (schema level). Facing to this issue, the OLAP analysis oriented approach consists of transforming SW data into OLAP cube via ETL processes. In this way, OLAP analysis can be carried out over extracted SW data through existing analysis tools. In the following, we will discuss about several works using this approach.

(Romero et Abelló, 2007) propose a semi-automatic approach to define an OLAP data warehouse from a single domain ontology. The resulting data warehouse could potentially integrate heterogeneous web sources while following a traditional OLAP data model. This approach enables OLAP analysis to be carried out over extracted SW data. However, valuable information can be found in several domain ontologies in a real-world application. Since the approach proposed in (Romero et Abelló, 2007) is based on a single domain ontology, it does not provide solution for reconciling overlapping concepts in different domain ontologies. (Nebot *et al.*, 2009) propose a framework to define semi-structured data warehouse from multiple domain ontologies. This data warehouse, called *Semantic Data Warehouse* (SDW), uses ontology mappings in order to manage domain overlappings. Coherent instances from different domain ontologies are derived and then assembled to semi-automatically generate a targeted OLAP cube.

These works focus on extracting, transforming and loading SW data into OLAP cubes so that decision-makers can directly carry out OLAP analysis. The main advantage is the possibility of reusing existing OLAP tools while analyzing transformed SW data in OLAP cube. However, storing SW data into a relatively static local data warehouse goes against the highly dynamic nature of web-published information. Moreover, the ETL process is not yet totally automatic but quite time-consuming (Romero et Abelló, 2007 ; Nebot *et al.*, 2009 ; Pardillo et Mazon, 2011). From a user's perspective, i.e. requiring high data freshness but not necessarily continuous querying (Pedersen, Castellanos, et Dayal, 2015), semi-automatically or manually built local SW data warehouse can hardly react to changes in data sources in real-time. As a result, the consistency between warehoused data and data in online sources is hard to be maintained. The quality of decision would be low if decision-makers analyze obsolete data in an agile business context.

3.2 Multidimensional modeling oriented approach

To overcome the drawbacks of previous approaches, the other research axis consists of carrying out multidimensional analysis directly over SW data without ETL processes. Most of the current frameworks are based on the RDF Data Cube vocabulary (QB), a core vocabulary proposed by W3C aiming to publish statistical and multidimensional datasets in the RDF standard. Directly effectuating OLAP analysis over QB-based model seems to be more efficient because no more ETL process is required. But the principle question is that OLAP analysis requires a complex model of data cubes containing facts, dimensions, multiple hierarchies and levels (Ibragimov *et al.*, 2014). Even thought QB allows representing hierarchical relationships between dimension instances via skos: narrower[7], it does not provide mechanism to represent multiple levels on a dimension and the relationships between levels at schema level. (Etcheverry et R. A. Vaisman, 2012).

Facing to this issue, (Kämpgen, O'Riain, et Harth, 2012) define an extension of QB model in order to represent statistical data in a multidimensional model. They illustrate how to carry out OLAP analysis over data published in QB by using the SPARQL[8] query language. However, their solution does not support dimensions with multiple hierarchies. Consequently, (Etcheverry et A. A. Vaisman, 2012) introduce a new multidimensional modeling language called Open Cube (OC), which supports multiple hierarchies in a dimension. Implementation of OLAP

[7] http://www.w3.org/2009/08/skos-reference/skos.html

[8] http://www.w3.org/TR/sparql11-overview/

operators through SPARQL queries are also presented in this work. However, OC is a specific modeling language, hence data already published in QB (which is standardized), cannot be reused by OC. To overcome this issue, (Etcheverry et R. A. Vaisman, 2012) introduce the QB4OLAP vocabulary. QB4OLAP extends and remains compatible with QB to support multidimensional modeling of SW data. In (Etcheverry, Vaisman, et Zimányi, 2014), an extension of QB4OLAP is proposed. It supports dimension with multiple hierarchies and it takes into account cardinalities between level members. Mechanisms to transform an existent relational data warehouse into QB4OLAP schema have also been presented in (Etcheverry, Vaisman, et Zimányi, 2014). The bi-directional compatibility between QB and QB4OLAP makes querying QB4OLAP with SPARQL possible, but issues about carrying out OLAP analysis in QB4OLAP model rather than simply querying still remains to be discussed. (Saad, Teste, et Trojahn, 2013) propose a conceptual multidimensional model based on QB which supports multi-facts, multi-dimensions and multi-hierarchies with different types (non-covering hierarchy). They also show how to implement OLAP operators via SPARQL queries with the proposed multidimensional model. To the best of our knowledge, (Saad, Teste, et Trojahn, 2013) were the first to address OLAP operators implementation through SPARQL queries in a complete multidimensional data model.

The multidimensional modeling oriented approach overcomes the problems of non-automaticity of ETL process: it provides compatible multidimensional modeling solutions for OLAP analyses over SW data. However, one fundamental principle of BI area, i.e., the materialization of data, is not fully taken into account by this approach. Most of the time, large datasets of SW data are queried on-the-fly, hence the efficiency of OLAP analysis using QB-like model is quite low (Kämpgen et Harth, 2013). Moreover, the quality of datasets varies from one to another; raw SW data without cleansing process may bring false information to decision-makers.

In the following table, we provide a summarized comparison of all listed works belonging to the two approaches.

Table 1. Summarized Comparison

OLAP analysis approach	Advantages : Reuse of existing OLAP technologies and tools Disadvantages : ETL process non-automatic				
	Heterogeneous data sources	OLAP analysis	Multiple ontologies	Automatic cube generation	
(Romero et Abelló, 2007)	√	√			
(Nebot *et al.*, 2009)	√	√	√		
Multidimensional modeling approach	Advantages : Without need of ETL process Disadvantages : low efficiency of analysis				
	Multiple levels	Multiple hierarchies	Reuse standard	Querying	OLAP operators
QB			N/A	√	
(Kämpgen, O'Riain, et	√		√	√	

Harth, 2012)					
(Etcheverry et A. A. Vaisman, 2012)	√	√		√	√
(Etcheverry et R. A. Vaisman, 2012)	√	√	√		
(Etcheverry, Vaisman, et Zimányi, 2014)	√	√	√	√	
(Saad, Teste, et Trojahn, 2013)	√	√	√	√	√

4. Contextualization of business analysis

Other than being used as data sources for analysis, SW data can also be exploited as complementary information to explain the context of business analysis. For instance, the web-published news talking about steady high temperature in a region could explain the increasing sales of air-conditioners. The combination of external SW data with factual data in an OLAP data warehouse provides decision-makers with multiple views over their business activities. Identifying relevant SW data to contextualize business analysis is a promising way to build decision support systems of the next generation, yet the contextualization of OLAP analysis is achieved mainly through text mining and information retrieval technologies (Perez *et al.*, 2008). As far as we know, no research has fully taken advantage of SW technologies to provide context for analysis. In this section, we briefly present existing techniques for OLAP analysis contextualization, wishing to provide inspiration for future research combining BI with SW.

Contextualization of business analysis can be achieved by retrieving relevant information stored in different systems. (Priebe, 2004) present a prototype permitting to associate relevant documents in content management system with predefined OLAP reports in OLAP system. Through the prototype he envisions different components of an enterprise portal that should share user’s context in order to present separately stored but related information together. A formal approach permitting to communicate users’ analysis context is presented in (Priebe, 2005). By using mechanisms of meta-searching over heterogeneous metadata, related factual and non factual data can be presented together so as to explain the context of business analysis. The meta-searching is based on metadata enriched with ontological concept mappings. The ontological concept mapping permits to associate the same concept in heterogeneous data sources to the same metadata. This provides a solution for handling the heterogeneity of data in different sources.

The approach proposed by (Priebe, 2005) allows a component of enterprise portal to communicate current user’s task with other components, so that all components in a portal could display various information related to a given analysis context. The quality of contextualization mainly depends on information embedded in metadata. However, if decision-makers could freely express their analysis context, the contextualization process would be more flexible and more adaptable to users’ needs. To this end, (Manuel Pérez-Martínez *et al.*, 2008) present an architecture of data warehouse contextualized with documents. By integrating relevant document segments in OLAP cube, this contextualized data warehouse provides decision-makers with information ranked on the basis of relevance to current analysis context. While analyzing, decision-makers can visualize related document segments along with factual data in OLAP cube. The work of (Manuel Pérez-Martínez *et al.*, 2008) differs from (Priebe, 2005) mainly because (Manuel Pérez-Martínez *et al.*, 2008) permit decision-makers to express their own analysis context.

Another way to contextualize business analysis is to retrieve related information on Internet. (Roy *et al.*, 2005) present an approach to associate relevant unstructured data from web with factual data in data warehouse. Firstly, a set of keywords is obtained by exploiting SQL query results. Then, the set of keywords is augmented with more terms retrieved by following the foreign-keys pointers between tables in the data warehouse. At last, the augmented set of keywords is used to retrieve web information via a keyword-based search engine (e.g., Google), so that the analysis context can be explained by the returned search results. This approach is not based on additional semantic information other than factual data in the relational database. Of course, SW techniques (e.g., ontologies) would surely increase keyword retrieval quality. (Liu, Xin, et Alon Y, 2006) propose a mechanism to extract keywords from structured query itself without the need of query execution: instead of obtaining information from query's result, they exploit information embedded in the query. A query is transformed in a set of keywords by removing distractive and unrelated information. The extracted keywords are then used for keyword-based search in a search engine so as to provide analysis context. This is a more generic approach comparing to (Roy *et al.*, 2005), because all types of structured query (SQL query, XML query etc.) are supported by (Liu, Xin, et Alon Y, 2006). Furthermore, in this work we can find further discussion about the benefits of combining keyword extraction with domain knowledge. However, this discussion is very imprecise, a concrete integration strategy of keyword extraction with SW technologies is still missing in this work. What's more, all above-mentioned works are based on traditional IR technologies. We believe new IR research results would certainly improve the efficiency of contextualization process. For instance, *Semantic IR* can be used to exploit semantic meanings embedded in web resources (Fernández *et al.*, 2011). Thus, if the contextualization process has been built on *Semantic IR*, the returned results would be more accurate and more complete.

(Castellanos *et al.*, 2010) and (Castellanos *et al.*, 2012) propose a framework along with a prototype allowing identifying external events in streaming data that would potentially affect the business operations. Based on text-mining techniques, this framework permits to extract and correlate textual information from internal and external data sources. In this way, newly generated web information is constantly associated with related internal information, which provides decision-makers an up-to-date context for their decisions. The following table presents a synthetic view of aforementioned work.

Table 2. Summarized Comparison

	Heterogeneous data sources	Storage of retrieved context	Up-to-date information	OLAP analysis	Ontology based
(Priebe, 2004) and (Priebe, 2005)	√	√			√
(Manuel Pérez-Martínez *et al.*, 2008)	√	√		√	
(Roy *et al.*, 2005)	√		√		
(Liu, Xin, et Alon Y, 2006)	√		√		
(Castellanos *et al.*, 2010) and (Castellanos *et al.*, 2012)	√	√	√		

5. Future research direction

Various challenges need to be overcome before a complete and efficient combination of BI with SW. For instance, concerning SW data storage (Niinimäki et Niemi, 2009 ; Deliège et Pedersen, 2010 ; Nebot et Berlanga, 2012) and data aggregation reasoning (Calvanese *et al.*, 2008 ; Thorne et Calvanese, 2009). In this section, we mainly focus on two specific issues: data materialization and SW data integration, because few proposals related to these issues are made to fully take advantage of both BI and SW domains.

5.1. Data Materialization

One of the fundamental principles of data warehouse in the BI area is the materialization of data. Researches belong to the approach oriented OLAP analysis consist in a full materialization through ETL process at the price of losing the data freshness. On the other hand, multidimensional modeling oriented approach ignores data materialization: SW data are extracted and queried on-the-fly, which brings about problems in terms of querying efficiency and data quality. To overcome the above-mentioned problems, a promising future research direction consists in partially materializing SW in data warehouse. This partial materialization should be performed at two levels: raw data and aggregated data.

Raw data refer to initial web-published data that are not yet subjected to analysis. At raw data level, not all data but only some relatively stable SW data should be maintained in data warehouse. By mentioning stable data, we refer to read-only or read-mostly data with little change over time, such as country's name for geographical data. For insert heavy datasets, only data in very common analysis path should be materialized. Moreover, only relevant data in large online datasets should be materialized. Avoid warehousing irrelevant data requires a precise and efficient data acquisition process. Extensions of classical ETL technologies should be defined to include new data acquisition rules. (Dayal *et al.*, 2009) point out that inspirations can be found from *Rule Learning* (Stephen, 1999) and *Hidden Markov Models* (Freitag et McCallum, 2000).

In the context of data warehouse, aggregated data refer to pre-summarized information that aims at accelerating analyses over regularly used data. Traditional OLAP tools already allow materializing aggregated data at different granularity levels. However, with the arrival of SW data in traditional OLAP data cube, the materialization of semantic graph data (e.g., RDF) does not always increase the efficiency of analysis if we follow classical aggregation rules. What's worse, analysis becomes sometimes less efficient in certain conditions with traditional aggregation functions (Kämpgen et Harth, 2013). New aggregation rules and functions need to be defined to support materialization of aggregated graph data in an efficient way. Inspiration can be found within *Query Shortcuts* technologies. More specifically, we can consider the materialized aggregated data as a set of shortcuts between the fact and certain disjunctive hierarchical levels in a graph model. Thus, based on the proposed algorithms in (Dritsou *et al.*, 2011), we can decide which shortcuts should be materialized in order to get the best trade-off between querying efficiency and optimal volume of data storage.

5.2. Automatic integration of SW data in OLAP cube

The common method to deal with unstructured (or less structured) data in OLAP data cube is to create data mappings through ontology. Most existing approaches assume that such ontology is easily built if not provided beforehand. In fact, in many cases finding an appropriate ontology for a specific domain is not a trivial work. On the other hand, building ontology from scratch is extremely complicated and thus not recommended. Therefore, automatically creating mappings between heterogeneous data with and without existing ontology is one of the future research challenges. Solutions for this issue can be found within the SW domain, especially the ones based on ontology alignment (Euzenat, 2013). For instance, a primitive data integration process can be manually defined with the help of semantic annotation and ontology mapping (Skoutas et Simitsis, 2007). This preliminary and manually-defined process could simply the automatic definition of future data integration process both in schema-level and instance-level (Rahm et Bernstein, 2001).

6. Conclusion

This paper provides an up-to-date overview of researches aiming to enhance OLAP analysis in the BI field with SW technologies. We can notice traditional OLAP can hardly deal with data coming from heterogeneous and external sources in open-world analysis scenarios. SW technologies come to rescue as they have been conceived to build semantic spaces over online information so that both humans and machines can get the correct semantic meaning of web published data. Enhancing OLAP analysis with SW technologies is a promising way to include external and heterogeneous information in traditional analysis process.

We discussed recent research results according to these approaches: (a) OLAP-analyses oriented approach which uses ETL process to integrate SW data in traditional OLAP data warehouses; (b) multidimensional modeling oriented approach which aims to define an appropriate multidimensional data model supporting direct OLAP analyses over RDF data collections. We concluded that SW technologies can indeed bring powerful tools to OLAP analysis, and OLAP can be used to efficiently analyze SW data. However, future research efforts are still needed to achieve a complete combination of OLAP with SW.

We envision a new data warehouse approach, which may be contextualized with SW data. This approach provides a promising solution for the restitution of both factual data and SW data during an analysis process. A number of researches have involved the contextualization of business analysis with external information by means of text mining or information retrieval. We believe that SW technologies will surely reinforce the ability of contextualization by providing semantically annotated information over web-published data.

Some directions for future research are outlined to make the best use of the two domains. We believe that fundamental principle of BI, such as data materialization, could improve efficiency and quality of analysis over SW data, while SW technologies, such as semantic annotations and ontology alignments, could provide theoretical and algorithmic basis for data warehouse evolution.

References

Abelló, Alberto, Jérôme Darmont, Lorena Etcheverry, Matteo Golfarelli, Jose-Norberto Mazón, Felix Naumann, Torben Pedersen, Stefano Bach Rizzi, Juan Trujillo, Panos Vassiliadis, et Gottfried Vossen. 2013. « Fusion Cubes: Towards Self-Service Business Intelligence ». *Int. J. Data Warehous. Min.* Vol. 9, n°2, p. 66‑88.

Abelló, Alberto, Oscar Romero, Torben Bach Pedersen, Rafael Berlanga, Victoria Nebot, Maria Jose Aramburu, et Alkis Simitsis. 2015. « Using Semantic Web Technologies for Exploratory OLAP: A Survey ». *IEEE Trans. Knowl. Data Eng.* Vol. 27, n°2, p. 571‑588.

Antoniou, Grigoris, et Frank van Harmelen. 2004. « Web Ontology Language: OWL ». Dans : *Handb. Ontol.* Berlin, Heidelberg : Springer Berlin Heidelberg, p. 67‑92.

Aufaure, Marie-Aude, et Raja Chiky. 2014. « From Business Intelligence to Semantic Data Stream Management ». Dans : *Adv. Concept. Model.* Cham : Springer International Publishing, p. 85‑93. ISBN : 978-3-319-12255-7, 978-3-319-12256-4.

Berlanga, Rafael, María José Aramburu, Dolores M. Llidó, et Lisette García-Moya. 2014. « Towards a Semantic Data Infrastructure for Social Business Intelligence ». Dans : *New Trends Databases Inf. Syst.* Cham : Springer International Publishing, p. 319‑327.

Berners-Lee, Tim, James Hendler, et Ora Lassila. 2001. « The semantic web ». *Sci. Am.* Vol. 284, n°5, p. 28‑37.

Calvanese, Diego, Evgeny Kharlamov, Werner Nutt, et Camilo Thorne. 2008. « Aggregate queries over ontologies ». Dans : *Proc. 2nd Int. Workshop Ontol. Inf. Syst. Semantic Web*. New York : ACM Press, p. 97‑104. ISBN : 9781605582559.

Castellanos, Malu, Chetan Gupta, Song Wang, Umeshwar Dayal, et Miguel Durazo. 2012. « A platform for situational awareness in operational BI ». *Decis. Support Syst.* Vol. 52, n°4, p. 869‑883.

Castellanos, Malu, Song Wang, Umeshwar Dayal, et Chetan Gupta. 2010. « SIE-OBI: a streaming information extraction platform for operational business intelligence ». Dans : *Proc. ACM SIGMOD Int. Conf. Manag. Data*. Indianapolis, Indiana, USA : ACM Press, p. 1105‑1110. ISBN : 9781450300322.

Chaudhuri, Surajit, Umeshwar Dayal, et Vivek Narasayya. 2011. « An overview of business intelligence technology ». *Commun. ACM*. Vol. 54, n°8, p. 88‑98.

Chen, Hsinchun, Roger H. L. Chiang, et Veda C. Storey. 2012. « Business intelligence and analytics: from big data to big impact ». *MIS Q.* Vol. 36, n°4, p. 1165‑1188.

Codd, E. F., S. B. Codd, et C. T. Salley. 1993. *Providing OLAP (On-line Analytical Processing) to User-analysts: An IT Mandate*. Codd & Associates,

Dayal, Umeshwar, Malu Castellanos, Alkis Simitsis, et Kevin Wilkinson. 2009. « Data integration flows for business intelligence ». Dans : *Extending Database Technol. Adv. Database Technol.* New York, NY, USA : ACM Press, p. 1‑11.

Deliège, François, et Torben Bach Pedersen. 2010. « Position list word aligned hybrid: optimizing space and performance for compressed bitmaps ». Dans : *Proc. 13th Int. Conf. Extending Database Technol.* New York : ACM Press, p. 228‑239.

Dritsou, Vicky, Panos Constantopoulos, Antonios Deligiannakis, et Yannis Kotidis. 2011. « Optimizing Query Shortcuts in RDF Databases ». Dans : *Semanic Web Res. Appl.* Berlin, Heidelberg : Springer Berlin Heidelberg, p. 77‑92.

Etcheverry, Lorena, et Alejandro A. Vaisman. 2012. « Enhancing OLAP Analysis with Web Cubes ». Dans : *Semantic Web Res. Appl.* Berlin, Heidelberg : Springer Berlin Heidelberg, p. 469‑483.

Etcheverry, Lorena, Alejandro Vaisman, et Esteban Zimányi. 2014. « Modeling and Querying Data Warehouses on the Semantic Web Using QB4OLAP ». Dans : *Data Warehous. Knowl. Discov.* Cham : Springer, p. 45‑56.

Etcheverry, Lorena, et Ro A Vaisman. 2012. *QB4OLAP: A New Vocabulary for OLAP Cubes on the Semantic Web*.

Euzenat, Jerome. 2013. *Ontology matching*. 2nd edition. New York : Springer,

Fernández, Miriam, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, et Enrico Motta. 2011. « Semantically enhanced Information Retrieval: An ontology-based approach ». *Web Semant. Sci. Serv. Agents World Wide Web*. Vol. 9, n°4, p. 434‑452.

Finin, T., J. Mayfield, A. Joshi, R.S. Cost, et C. Fink. 2005. « Information Retrieval and the Semantic Web ». Dans : *Proc. 38th Hawaii Int. Conf. Syst. Sci.* IEEE, p. 1‑10.

Freitag, Dayne, et Andrew McCallum. 2000. « Information Extraction with HMM Structures Learned by Stochastic Optimization. » Dans : *Innovative Applications of Artificial Intelligence Conferences*. Austin, Texas : AAAI Press / The MIT Press, p. 584‑589.

Harinarayan, Venky, Anand Rajaraman, et Jeffrey D. Ullman. 1996. « Implementing data cubes efficiently ». *ACM SIGMOD Rec.* Vol. 25, n°2, p. 205‑216.

Horrocks, Ian, Peter F. Patel-Schneider, et Frank van Harmelen. 2003. « From SHIQ and RDF to OWL: the making of a Web Ontology Language ». *Web Semant. Sci. Serv. Agents World Wide Web*. Vol. 1, n°1, p. 7‑26.

Ibragimov, Dilshod, Katja Hose, Torben Bach Pedersen, et Esteban Zimányi. 2014. « Towards Exploratory OLAP over Linked Open Data–A Case Study ». Dans : *BRITE*. HangZhou : p. 1‑18.

Inmon, William H. 1996. *Building the data warehouse*. 2nd ed. New York : Wiley Computer Pub, 401 p. ISBN : 0471141615.

Kämpgen, Benedikt, et Andreas Harth. 2013. « No Size Fits All – Running the Star Schema Benchmark with SPARQL and RDF Aggregate Views ». Dans : *Semantic Web Semant. Big Data*. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 290‑304.

Kämpgen, Benedikt, et Andreas Harth. 2011. « Transforming statistical linked data for use in OLAP systems ». ACM Press, p. 33‑40. ISBN : 9781450306218.

Kämpgen, Benedikt, Sean O'Riain, et Andreas Harth. 2012. « Interacting with statistical linked data via olap operations ». Dans : *Int. Workshop Linked APIs Semantic Web LAPIS 2012. 9th Extended Semantic Web Conference*. Heraklion, Greece : Citeseer, p. 36‑49.

Kimball, Ralph, éd. 1998. *The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses*. New York : Wiley.

Kimball, Ralph. 1996. *The data warehouse toolkit: practical techniques for building dimensional data warehouses*. New York : John Wiley & Sons.

Liu, Jing, Dong Xin, et Halevy Alon Y. 2006. « Answering Structured Queries on Unstructured Data ». Dans : *Web and Databases (WebDB)*. Chicago, Illinois : p. 20‑25.

Manuel Pérez-Martínez, Juan, Rafael Berlanga-Llavori, María José Aramburu-Cabo, et Torben Bach Pedersen. 2008. « Contextualizing data warehouses with documents ». *Decis. Support Syst.* Vol. 45, n°1, p. 77‑94.

McBride, Brian. 2004. « The Resource Description Framework (RDF) and its Vocabulary Description Language RDFS ». Dans : *Handb. Ontol.* Berlin, Heidelberg : Springer Berlin Heidelberg, p. 51‑65. ISBN : 978-3-662-11957-0.

Nebot, Victoria, et Rafael Berlanga. 2012. « Building data warehouses with semantic web data ». *Decis. Support Syst.* Vol. 52, n°4, p. 853‑868.

Nebot, Victoria, Rafael Berlanga, Juan Manuel Pérez, María José Aramburu, et Torben Bach Pedersen. 2009. « Multidimensional Integrated Ontologies: A Framework for Designing Semantic Data Warehouses ». Dans : *J. Data Semant. XIII*. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 1‑36.

Niinimäki, Marko, et Tapio Niemi. 2009. « An ETL Process for OLAP Using RDF/OWL Ontologies ». Dans : *J. Data Semant. XIII*. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 97‑119.

Pardillo, Jesus, et Jose Norberto Mazon. 2011. « Using Ontologies for the Design of Data Warehouses ». *Int. J. Database Manag. Syst.* Vol. 3, n°2, p. 73‑87.

Pedersen, Torben Bach, Malu Castellanos, et Umesh Dayal. 2015. « Report on the Seventh International Workshop on Business Intelligence for the Real Time Enterprise (BIRTE 2013) ». *ACM SIGMOD Rec.* Vol. 43, n°4, p. 55‑58.

Perez, J.M., R. Berlanga, M.J. Aramburu, et T.B. Pedersen. 2008. « Integrating Data Warehouses with Web Data: A Survey ». *IEEE Trans. Knowl. Data Eng.* Vol. 20, n°7, p. 940‑955.

Priebe, Torsten. 2005. « Building Integrative Enterprise Knowledge Portals with Semantic Web Technologies. » Dans : *Intell. Learn. Infrastruct. Knowl. Intensive Organ. Semantic Web Perspect.* IGI Global, p. 146‑188. ISBN : 9781591405030, 9781591405054.

Priebe, Torsten. 2004. « INWISS–Integrative Enterprise Knowledge Portal ». Dans : *Demonstr. 3rd Int. Semantic Web Conf. ISWC 2004*. Hiroshima, Japan : p. 33‑36.

Rahm, Erhard, et Philip A. Bernstein. 2001. « A survey of approaches to automatic schema matching ». *VLDB J.* Vol. 10, n°4, p. 334‑350.

Ravat, Franck, Olivier Teste, Ronan Tournier, et Gilles Zurfluh. 2008. « Algebraic and Graphic Languages for OLAP Manipulations »: *Int. J. Data Warehous. Min.* Vol. 4, n°1, p. 17‑46.

Romero, Oscar, et Alberto Abelló. 2007. « Automating multidimensional design from ontologies ». Dans : *Int. Workshop Data Warehous. OLAP*. ACM Press, p. 1‑8.

Roy, Prasan, Mukesh Mohania, Bhuvan Bamba, et Shree Raman. 2005. « Towards automatic association of relevant unstructured content with structured query results ». Dans : *Int. Conf. Inf. Knowl. Manag.* ACM Press, p. 405‑412. ISBN : 1595931406.

Saad, Rafik, Olivier Teste, et Cássia Trojahn. 2013. « OLAP Manipulations on RDF Data following a Constellation Model ». Dans : *Int. Semantic Web Conf. ISWC2013*. Sydney :

Skoutas, Dimitrios, et Alkis Simitsis. 2007. « Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data »: *Int. J. Semantic Web Inf. Syst.* Vol. 3, n°4, p. 1‑24.

Stephen, Soderland. 1999. « Learning Information Extraction Rules for Semi-Structured and Free Text ». *Mach. Learn. - Spec. Issue Nat. Lang. Learn.* Vol. 34, n°1-3, p. 233 ‑ 272.

Thi, A, et Binh Thanh Nguyen. 2008. « A Semantic approach towards CWM-based ETL processes ». *Proc. -Semant.* Vol. 8, p. 58‑66.

Thorne, Camilo, et Diego Calvanese. 2009. « Controlled Aggregate Tree Shaped Questions over Ontologies ». Dans : *Flex. Query Answering Syst.* Springer Berlin, p. 394‑405.

Trujillo, Juan, et Alejandro Maté. 2012. « Business Intelligence 2.0: A General Overview ». Dans : *Bus. Intell.* Berlin, Heidelberg : Springer Berlin Heidelberg, p. 98‑116.

Zorrilla, Jose-Norberto, Óscar, Irene, Florian Daniel, et Juan Trujillo, éd. 2012. *Business Intelligence Applications and the Web: Models, Systems and Technologies*. IGI Global, ISBN : 9781613500385, 9781613500392.