TOWARD AN ONTOLOGY BASED APPROACH FOR DATA WAREHOUSING
STATE OF THE ART AND PROPOSAL

ELHAJ ELAMIN¹ and JAMEL FEKI²
¹Sudan University of Science and Technology, Khartoum, Sudan
²Laboratory Mir@el, University of Sfax, Sfax, Tunisia
elhajelamin@hotmail.com; Jamel.feki@fsegs.Rnu.t

ABSTRACT

The design process of a data warehouse (DW) raises many problems and then is considered as a complex and tedious task. DW designers follow different levels concerning the design process; among these levels the conceptual level gains a significant care in the literature so far. Recently, using a hybrid approach conjugated to a semantic resource is considered as a successful way for designing a DW that meets the decision makers' requirements, the organization’s operational data source, and the design quality. Many works have been proposed, as well, different approaches or/and methods have been applied for designing a DW. However, still there is not yet a common approach that designers can follow to design a DW. This paper is twofold objectives: First, we conduct a comprehensive survey among proposed works in order to compare their methods based on a set of significant criteria. Secondly, considering the result of this study, we suggest an ontology-based hybrid approach for the DW design. The result is a framework consisting in three main steps: i) Building star schemas based on business-requirements; ii) Building star schemas from an operational data-source; and iii) Matching the two sets of schemas to produce a DW model compliant to user and data source simultaneously. In order to enhance these steps, each one uses an ontology as a knowledge representation to alleviate semantic issues.

Keywords: Decision Support System, Data Warehouse, Multidimensional model, Star schema, Semantic resource, Conceptual design.

1. INTRODUCTION

So far, the data warehouse (DW) design process raises many problems and then is considered as a complex and tedious task. This complexity in one hand appears as the absence of standard concepts and terminology between various stakeholders involved in the DW design process (i.e., uncommon terminological and semantic referential) and the lack of methodological framework and software tools those assist DW designers during the development process [18]. On the other hand, the complexity is due to three types of heterogeneity: syntactic, structural and semantic [1][2].

In the DW literature, the proposal to overcome these difficulties is the integration of the data sources into a unified repository (i.e., DW) to easily and correctly perform activities like querying and reporting for the support of strategic decision making processes [12].

The main data warehousing phases are: Requirements elicitation in terms of analytical analysis; conceptual design; logical design, and physical design [5]. Furthermore, there are many directions one can follow to design a DW. Indeed, some designers prefer to start from user requirements (top-down approaches) [11][13][16]; the motivation for the camp that uses this approach is that the requirement analysis phase is crucial to meet the user needs and expectations. However, the problem here is the risk to miss some pertinent analyses, simply because users have no skills in expressing their needs and are not familiar with this task. Another camp of DW designers prefers to start the design process from the data source (DS) data model which is the operational database, applying what is so called data-source driven or bottom-up approaches [7][6][21]. The main drawback of these approaches is the obtained DW model could be larger, in terms of structure, than the interests of decision makers [14]. A third camp of designers merge top-down and bottom-up approaches leading to a hybrid approach [1][3][9] which benefits from the features of the above two approaches; i.e., takes into account user requirements and the DS model; in addition they confront the two output models for validation purposes. Right now, and to the best of our knowledge, there are few attempts that fully automate this hybrid approach. Furthermore, the use of a semantic resource during all the steps of hybrid approaches still remains weak.

In fact, hybrid approaches can be classified into pure hybrid approaches and integration driven hybrid approaches. The former considers during the design process each of the DS and business goals separately,
 Whereas the later integrates a data source model driven approach with user requirements driven one [22] [45]. They may be achieved either sequentially: means the two stages executed in a predefined order; or even in parallel, when the two stages executed independently, the comparison and integration of schemas are performed later [45]. Figure 1 recapitulates the DW design approaches.

**FIGURE 1. DATA WAREHOUSE DESIGN APPROACHES [45]**

During the last decade, the concept of ontology has emerged as a promising key component in information systems. Ontology is a means for modeling the semantics of concepts used by various heterogeneous stakeholders in a well defined and unambiguous way [4]. Several researchers have introduced this concept in many computer field applications as an efficient means to overcome semantic difficulties mainly. The recorded success has encouraged the use of this ontology concept in Decision Support System (DSS). Using of ontology for designing a DW helps organizations to solve problems of semantic heterogeneity [9] between data sources and user requirements. Ontologies have been studied so far among the components of the DW design process; thus, ontological software tools have been proposed as well as some languages. However, there is not a clear vision of situations where ontologies may be applied and there is not still general agreement about how to do ontologies [9].

Among the objectives of this paper, is to make a survey about ontology methods for designing DWS. This paper is organized as follows: in section 2 we give an overview of related work, it contains a rich background about DW and ontology concepts in addition to the state of the art; as well we conduct a comparative study for DW design approaches. In section 3 we introduce our proposed framework for a hybrid and semi-automatic ontology-based approach for the DW design process. Finally, section 4 concludes the paper and highlights next steps.

2. RELATED WORK

In the literature of the DWSing field many works have been proposed, various approaches and methods have been used to assist the DW design process. In common to these methods, ontology had been considered as a key component to succeed this process. For the purpose of organizing this survey, we consider two axes: the first one represents issues related to the DW design, whereas the second one tackles ontologies in DWSing. In between we may merge the two axes as necessary.

2.1 DATA WAREHOUSING APPROACHES

Functionally, the DWSing process consists of three phases: i) Extracting data from operational sources, ii) Organizing and integrating data consistently into the DW, and iii) Accessing the integrated data in an efficient and flexible fashion [1] [6] [42] [54] [58] by decision makers.

In the literature review, DW designers follow different directions concerning the number of the design phases. Although the core phases are four [5], studies in [19] [4] proposed three steps relative to conceptual level, logical level and physical level, whereas the study in [25] follows the conventional core four steps. Another comprehensive study conducted by [26] defines eight steps to design a DW, these steps include: requirements analysis, analysis and reconciliation, conceptual design, workload refinement, logical design, data staging design, physical design and implementation.

Whatever the design approach, the design step concerning the conceptual level gains the significant care in the literature. Indeed, almost researchers and practitioners agree that the conceptual design is considered as a keystone on which depends the success or the failure of the DW project [6] [39] [27] [53] [58] [59]. Furthermore, the failure in DW projects is usually returned to the poorness and inadequacy of conceptual design methods [39] [41]. However, for our knowledge, there is no standard methodology designers can follow to achieve this conceptual design phase.

2.1.1 Conventional DW Design Approaches

Generally, the conceptual design phase can be accomplished either by starting from the data source (DS) data model or from user requirements, each of these methods has advantages and disadvantages. In recent years, the research trend follows a hybrid approach which provides solutions to the problems encountered in the former approaches; in this sense the study in [49] follows this approach where the authors analyze the requirements stated by decisional users in terms of SQL queries and, in parallel, they analyze the operational relational DS expected to feed the DW with data. Other works following the hybrid approach are [1]
[4] [16] [35] [45]; however, almost all works don't suggest a well founded solution for heterogeneity issues related to the contents of various sources.

A conceptual model dedicated for DWs baptized DFM (Dimensional Fact Model) was proposed by Golfarelli [6]; in this DFM model the representation of the decisional world (i.e., reality) is called dimensional schema. The basic concepts of the DFM are the fact, dimension and hierarchy.

Concerning the logical DW design, this step involves the definition of structures that enable an efficient access to information [40]; as it is less important than the conceptual design step, few works have touched this level; as an example, the studies in [19] [40] proposed an environment for generating logical schema from a conceptual schema; another attempt for DW logical design has been made by [44] where the authors developed a DW design methodology from XML sources.

Finally, physical design of DW could be considered as a weakly treated phase through the literature [15] [42]; the study in [48] proposed a guidance for experts by providing best practice recommendations that designers can apply when designing a physical DW model. Another study dedicated to the physical design has been tackled by [25] where the authors presented physical modeling techniques; nevertheless, their approach encompasses as well conceptual and logical modeling.

### 2.1.2 Ontology-Based DW Design Approaches

The second axis we elected to organize this survey reveals how the ontology was used in various researches that applied the hybrid approach. As ontologies may vary in different aspects, these may have: content, structure, implementation and a conceptual scope [30]. For instance, the study in [1] transforms each data source into ontology format, to do so authors used RDBtoOnto tool to convert a relational database source into ontology; XML and text sources are converted by JXML2OWL and OntoLT tools respectively; they used the global ontology as input which has been build from local ontologies by Protege tool; the requirements represented in logical format by logical converter so as to map it with global ontology; after that, they can extract automatically facts and dimensions. Once again, in [1] the authors propose an ontology based hybrid approach to derive a DW multidimensional schema. However their suggested approach relying on transforming the data source model into an ontology reduces the genericity of their proposal, instead it is better to reuse a general ontology and involve the decision maker during the design process. This will make the DW design process more rapid and less costly.

In [2] user requirements and data sources are represented using ontology as in [1], but the difference between the two studies is that in [2] the authors proposed algorithms to generate the multidimensional elements (i.e., fact, dimension…). The authors in [4] defined an ontology based database making user requirements consistent within the DW structure; they proposed a requirement model following a goal driven approach; the specification of these requirements produces graphs which represent goals and their logical relationships in AND-OR graph format. The data sources in this study considered as ontology based database (OBDB) follow the architecture of OntoDB. Previous works dealing with reconciling user requirements with data sources [16] [38] follow a method in which a multidimensional model is obtained from user requirements, then, they enforce correctness between the obtained multidimensional model and data sources. In our knowledge, works that construct star schemas from business requirements and from operational data sources simultaneously have a major lack: they go through a transformation step for building an ontology; this step makes heavier and costly the DW design process.

In the same direction, in order to improve the DW design process, the study in [39] proposed a conceptual model for DWsing based on ontology; their method consists of two steps namely domain modeling and property modeling; domain modeling serves to describe user needs by a set of themes, each of these themes can be depicted by multiple dimensions, which in turn can be further divided into a variety of classification; whereas property modeling serves to solve the heterogeneity problem by using dimension coming from different platform.

The authors in [3] focused on a conceptual method for the DW design from data residing in a type of DB called ontology based DB. Another research [31] employed an ontology method to facilitate the conceptual design of the DW; the authors used a graph based representation to deal with both structured and semi-structured data. In further works [35] the authors proposed a hybrid approach in accordance with the MDA paradigm (Model Driven Architecture), for which they developed a conceptual model qualified as a PIM (Platform Independent Model) starting from users requirements model; this model is then reconciled with the data sources that will feed up the DW under construction.

### 2.2 COMPARING DW DESIGN APPROACHES

In order to reveal the pros and cons of various DW design approaches, we suggest comparing these approaches. To do so, first we adopt the well-known classification: Top-down, Bottom-Up and Hybrid. For this comparison we suggest the following criteria:

- Automation: either the approach is automatic, semi-automatic or manual.
- Use of semantic resource by the approach.
- Input of the approach: this depends on the type of the approach and the representing method of semantic resource, if it had been used.
- Output of the approach: whether it is a DW conceptual schema or logical schema.

Table 1 explores the general criteria for DW design approaches. There are few studies that tackle the comparison between various methodologies for DW designing [2] [15] [22] [36] [42] [57]. However, these studies do not cover most issues related to the design process. In [2], authors compare only the ontology based approaches that represent data sources in ontology; as few papers gain this feature, the study restricted only to these papers. Another study [15] tackles the comparison of DW design approaches where the authors proposed an object oriented framework using UML during the design process, so their comparison study were based on this background; their used criteria did not include either the input to the designing process nor the output model, which we considered important. Another drawback of this study is that it is not comprehensive; authors concentrate on approaches dealing with UML in the design process, as well as tools. Authors in [22] make a survey containing a comparison in form of graphical view; although the survey seems to be comprehensive, but it is difficult for the reader to obtain information from this graph view. In [36] [37] the phase of requirements analysis is considered so far; authors in [36] survey the journey from user requirements to conceptual design phase so that the comparison focuses on the types of user requirements techniques; whereas [37] tackles in details the comparison between mixed driven approaches. Recently, survey studies had been conducted [42] [46]; authors in [42] compares the methodologies for the MD design of the DW using a systematic mapping as research methodology. Their methodology consist of eight steps applied to the target approaches, the results obtained from this study provide a rich calculated information that serves as a source for readers to know where the trend of research in this area is going now; but, their comparison does not encompass criteria about ways/methods used by these works to achieve the DW design. In [46], although the authors draw a readable way describing various DW design approaches; their comparison study was restricted to only semi automatic approaches.

<table>
<thead>
<tr>
<th>Work reference</th>
<th>Type of approach</th>
<th>Automatic</th>
<th>Use of semantic resource</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>[6]</td>
<td>Yes</td>
<td>Semi</td>
<td>No</td>
<td>Conceptual or logical schemas (RDB)</td>
<td>Dimensional Fact model</td>
</tr>
<tr>
<td>[7]</td>
<td>Yes</td>
<td>Semi</td>
<td>No</td>
<td>Relational OLTP DB</td>
<td>Snowflake schema</td>
</tr>
<tr>
<td>[8]</td>
<td>Yes</td>
<td>Semi</td>
<td>No</td>
<td>Enterprise schema of an operational DB</td>
<td>DW schema</td>
</tr>
<tr>
<td>[11]</td>
<td>Yes</td>
<td>Semi</td>
<td>No</td>
<td>UML class diagram representing BR</td>
<td>Physical MOLAP</td>
</tr>
<tr>
<td>[14]</td>
<td>Yes</td>
<td>Semi</td>
<td>Yes</td>
<td>Heterogeneous DS schemas</td>
<td>Decisional ontology</td>
</tr>
<tr>
<td>[17]</td>
<td>Yes</td>
<td>Semi</td>
<td>Yes</td>
<td>Ontology domain (DS + domain vocabularies)</td>
<td>Constellation schema</td>
</tr>
<tr>
<td>[21]</td>
<td>Yes</td>
<td>Semi</td>
<td>Yes</td>
<td>Ontology DS schema</td>
<td>Constellation schema</td>
</tr>
<tr>
<td>[23]</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Ontology DB schema</td>
<td>Star schemas</td>
</tr>
<tr>
<td>[24]</td>
<td>Yes</td>
<td>Semi</td>
<td>Yes</td>
<td>Natural language + ontology</td>
<td>Semantic DW</td>
</tr>
<tr>
<td>[28]</td>
<td>Yes</td>
<td>Semi</td>
<td>No</td>
<td>DTD + XML files</td>
<td>Fact schemas</td>
</tr>
<tr>
<td>[29]</td>
<td>Yes</td>
<td>Semi</td>
<td>No</td>
<td>Organizational model + decisional model + DS</td>
<td>Fact schemas</td>
</tr>
<tr>
<td>[32]</td>
<td>Yes</td>
<td>Semi</td>
<td>Yes</td>
<td>Ontology domain</td>
<td>Dimension hierarchies</td>
</tr>
<tr>
<td>[33]</td>
<td>Yes</td>
<td>Semi</td>
<td>No</td>
<td>Star schema from DS + star schema from BR</td>
<td>The conceptual design of</td>
</tr>
<tr>
<td>[35]</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Conceptual MD from BR + DS schema</td>
<td>Implementation code for</td>
</tr>
<tr>
<td>[38]</td>
<td>Yes</td>
<td>Semi</td>
<td>No</td>
<td>MD model from BR + MD normal forms from</td>
<td>DW conceptual schema</td>
</tr>
<tr>
<td>[41]</td>
<td>Yes</td>
<td>Semi</td>
<td>No</td>
<td>ER schema</td>
<td>ME/R schemas</td>
</tr>
<tr>
<td>[43]</td>
<td>Yes</td>
<td>Semi</td>
<td>Yes</td>
<td>XML Format + ontology</td>
<td>ETL + MD schema</td>
</tr>
<tr>
<td>[45]</td>
<td>Yes</td>
<td>Semi</td>
<td>No</td>
<td>BR + DS</td>
<td>Fact schemas</td>
</tr>
<tr>
<td>[47]</td>
<td>Yes</td>
<td>Semi</td>
<td>Yes</td>
<td>Entity relationship diagram</td>
<td>Star schema</td>
</tr>
<tr>
<td>[49]</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>BR + the integrated logical model of the DS</td>
<td>Constellation schema</td>
</tr>
<tr>
<td>[50]</td>
<td>Yes</td>
<td>Semi</td>
<td>Yes</td>
<td>Ontology domain</td>
<td>Relational DW</td>
</tr>
<tr>
<td>[52]</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Integrated DS + SQL queries</td>
<td>Constellation schema</td>
</tr>
<tr>
<td>[54]</td>
<td>Yes</td>
<td>Semi</td>
<td>No</td>
<td>Entity relationship diagram</td>
<td>Fact schemes</td>
</tr>
<tr>
<td>[55]</td>
<td>Yes</td>
<td>Semi</td>
<td>Yes</td>
<td>Entity relationship diagram</td>
<td>Star schema</td>
</tr>
<tr>
<td>[56]</td>
<td>Yes</td>
<td>Semi</td>
<td>Yes</td>
<td>Semantic data</td>
<td>Data Mart schema</td>
</tr>
</tbody>
</table>

**TABLE 1. EVALUATION OF DW DESIGN APPROACHES**

To summarize, all these works tackle the comparison of DW design methodologies; however, they based their study on special criteria; as well, they did not touch...
significant features such as: treating business requirements, treating data sources, use of semantic resource. In our survey, we make use of all these significant criteria; as well we refer to recent works. Table 1 and table 2 compare these recent works. This encourages us to propose a framework to fill the gaps in this domain.

Table 2 depicts works that follow the hybrid approach and use semantic resource. Our framework agrees with much of them in some criteria such as the semi automation and the input to the system, but the difference appear at the output column; none of these works generate star schemas from both data sources and user requirements.

From these studies, we can learn the following lessons:
- Using semantic resource is important in the DW design process to solve heterogeneities.
- The use of ontology requires skilled persons in both ontology and application domain.
- As much works studied so far build ontology based data base, the transformation process from data source/user requirements into ontology requires specific tool.
- Building ontology as part of DW design phases increase both time and coast of the design process.

To fill this gap we suggest using WordNet as a semantic resource to solve various heterogeneity issues, at the same time we avoid the mentioned problems.

3. PROPOSED APPROACH

Considering previous observations we suggest, in the remaining of this paper, a hybrid and semi-automatic ontology-based approach for the DW design process.

Being hybrid, on one hand business requirements will be elicited from decision-makers to build what we call business-requirements star schemas (BR-stars for short), and on the other hand, the structured DS in form of a DB model (e.g. relational) is recovered to build data-source star schemas (DS-Stars). Thus the proposed approach expects gaining the best features of top-down and bottom-up approaches: mainly by the involvement of the DW designer as well as decision makers during all its steps. The build of these two types of star schemas relies on a set of specific heuristics we define to identify multidimensional concepts (facts and their measures, dimensions with their attributes). Note that these heuristics are automated.

Furthermore, since it is hybrid the approach aims to build a validated conceptual DW model after reconciling BR-stars with DS-Stars. This step relies on a matching process and uses a general semantic resource to overcome the semantic conflicts between the two types of input schemas. For this step, we need quality metrics to assess the resemblance of schemas during this matching process. The goal behind the process of matching is to reveal the missing concepts between the two schemas; as well to deal with semantic heterogeneity.

<table>
<thead>
<tr>
<th>Schemas Matching &amp; Validation</th>
<th>Star Schemas Approval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Warehouse Designer</td>
<td></td>
</tr>
<tr>
<td>Multidimensional Schema</td>
<td></td>
</tr>
<tr>
<td>Construction</td>
<td></td>
</tr>
<tr>
<td>Semantic Resource</td>
<td></td>
</tr>
<tr>
<td>Structured Data Source</td>
<td></td>
</tr>
<tr>
<td>(Relational DB)</td>
<td></td>
</tr>
<tr>
<td>Montant_Vente</td>
<td></td>
</tr>
<tr>
<td>Qte_Vendue</td>
<td></td>
</tr>
<tr>
<td>蒙年_Vente</td>
<td></td>
</tr>
<tr>
<td>Approved Star Schemas</td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 2. OVERVIEW OF THE PROPOSED APPROACH

Among its main advantages, the proposed approach will not transform the data source or user requirements into ontology representation; instead it reuses a general semantic resource which is usually free open source as WordNet and, therefore reduces the time and cost of the conceptual design phase. In addition, it does not require skilled persons in the domain of ontology.

Figure 2 depicts that the DW designer is able to intervene mainly through the: i) Construction of multidimensional schemas steps, ii) Matching of these schemas, and iii) Approval of star schemas (i.e., validation). The feedback coming from the designer serves to design robust star schemas. We detail this proposed approach in the next subsections.

From table 1 and table 2, it is clear that our approach agree with other approaches in that: it is hybrid because it tackles both data source and business requirements; in addition, it is semi-automatic since it uses automated extraction rules and the DW designer contributes at different levels of the design process. Furthermore, it relies on a semantic resource in order to overcome the semantic conflicts in names of multidimensional elements.

Our framework consists of three main steps: i) Building business-requirements star schemas; ii) Building data-source star schemas; and iii) Matching the two sets of schemas to produce BR-and-DS compliant star schemas. Hereafter we explain each of these three steps.
3.1 BUILDING STAR SCHEMA FROM BUSINESS REQUIREMENTS

The first phase of our framework deals with constructing star schemas on behalf of business requirements; this phase encompasses three tasks: i) Business requirements elicitation; ii) requirements normalization; and iii) generation of MD schemas. The output of this phase is a set of BR-star schemas compliant to user requirements.

3.1.1 Business requirements elicitation

Requirements elicitation is a critical task in systems designing process; it can be defined as "a process through which the customers, buyers or users of a software system discover, reveal, articulate and understand their requirements" [51]. During this step stakeholders should avoid many difficulties associated with it as it is a social interaction activity. The tangible result from this step is a list of requirements; which can be expressed by natural language-like, as example we can represent requirements in commercial domain as follows:

"Analyze sales By customer name".
"Analyze sales By customer and Product".

Note that after the keyword Analyze we encounter the business activity to be analyzed (i.e., the Sales fact here), after By we find the analyzing criteria (i.e., dimensions or levels)... The Analyze may represent one of the aggregate functions (Sum, Avg, Min…) which will be selected later at the query time. The aim behind using natural language-like is its simplicity for expressing analytical requirements [34].

3.1.2 User requirements normalization

The main function of this task is to represent the gathered requirements according to a given format that facilitates avoiding the redundancy in the set of the elicited business requirements, the extraction of the multidimensional elements and then the star schema derivation. We intended to use a matrix representation of requirements. During this task the designer may modify/correct requirements structures to minimize data redundancy. The output should be a non-redundant set of well defined requirements.

3.1.3 Multidimensional schema generation

For this step, we define a set of heuristics (i.e., extraction rules) working on the matrix of requirements previously elaborated. These rules are for the extraction of the MD schema components (facts, dimensions, measures, etc.). This step generates a set of BR-star schemas compliant to the DFM formalism.

3.2 BUILDING STAR SCHEMA FROM DATA SOURCES

This is the main second step in our approach. It deals with constructing star schemas from a relational DB source. It is composed of the following tasks: i) Database model extraction; ii) Reverse engineering; and iii) Multidimensional schema generation. These tasks are described hereafter.

3.2.1 Database model extraction

The tables’ structures of the relational DB source are extracted from the DBMS repository by querying system-views. In addition the constraints are extracted. More precisely, primary key and foreign key constraints are vital for the next step that classifies tables into two classes namely Entity-table and Relationship-table

3.2.2 Reverse engineering on the database model

Let’s note that in DWsing entities are typically used to design dimensions whereas relationships are for building facts [54] [60]. So then, a reverse engineering task is required. It aims to know which DB table initially represents an entity of the real world (e.g., Product, Customer) and which table represents a relationship between entities (e.g., Order items). This classification into two classes is crucial because it highlights which table will be candidate to define a fact and which one may become a potential dimension.

![Figure 3: Ontology-Based Approach for DW Design](Image)

3.2.3 Multidimensional schema generation
This task aims to build MD schemas from the classified tables. Defining a set of appropriate heuristics is required for the automation of this task. These heuristics are to find out automatically the MD schemas elements (facts, dimensions, measures, etc.) and then to group them into star schemas. For instance, the rule to find a fact is like “Each Relationship-table transforms into a fact”; and then measure are extracted through the following rule “Each numeric attribute belonging to a relationship-table transformed into a fact F becomes a potential measure for F”…

3.3 MULTIDIMENSIONAL SCHEMAS MATCHING

This represents the final step of our approach. Activities in this step are: i) Comparing DS-Star schemas with BR-Star schemas; ii) Usage of a semantic resource; iii) Involvement of the DW designer; and iv) Generation of approved star schemas.

3.3.1 Comparing DS-Stars with BR-Star schemas

Star schemas issued from the DS are compared with those built from business requirements (BR). The objective here is to detect inconsistencies between elements. The inconsistency may come in different forms such as diversity in elements names, missing elements in one schema…. Naturally, this step should be optimized in the sense not a Cartesian product of matching is used; indeed, we match only schemas analyzing the same business activity, i.e., having identical, synonym fact names. To do so, we need to define semantic metrics to measure the resemblance between two schemas. This metric should also take into account the number of common dimensions for the two facts. The granularity of measures is a furthermore issue to solve, but not at the conceptual level since it could be differed to the implementation (i.e., physical) level.

3.3.2 Usage of semantic resource

A convincing matching process requires the usage of a semantic resource in order to identify whether a name of a given concept (i.e., fact or dimension) is semantically equivalent to another or not (e.g. part and piece are synonyms). In our framework, we use the free and open source WordNet as a general semantic resource.

3.3.3 Involvement of the DW designer

In our approach, the schemas matching process is semi-automatic; the output of this matching is a set of multidimensional components which may be incomplete for reflecting decision makers requirements; or have a form of inconsistency. The DW designer involvement in this activity helps to complete the correctness of the generated multidimensional schemas. The task of the DW designer is to confirm this output in means of semantic confirmation.

3.3.4 Generating approved star schemas

The final output of our framework is a set of approved star schemas; these schemas should satisfy the decision maker requirements and ensures that they could be loaded with data from the operational data source used in the design. Multidimensional constraints could be defined in order to enhance the schema validity e.g., avoid empty facts, empty dimensions… In addition, devising the designer in case of degenerated dimensions is possible.

4. CONCLUSION

In this paper we have conducted a comparative study involving several research works tackling the data warehouse (DW) design process. The DW represents the keystone in modern decision support systems (DSS). The DW design process is a complex, tedious and error-prone activity which lacks software tools that assist the DW designer and future users (i.e., decision makers) during the conceptual design steps. These steps, generally performed manually, require skilled persons with competences in several design specialties in classical information systems modeling in order to understand the data source structure when applying a bottom-up approach; in requirements elicitation when the approach design is top-down, the concepts and terminology of the application domain to solve semantic difficulties; and in multidimensional modeling.

Among the main conclusions of this study is there is a research trend to the usage of ontology during some of the conceptual design process; nevertheless, the efforts investigated so far in using the ontology concept focus on a single/isolated step of the DW design process. The objective of this paper was first to highlight this main lack and, secondly to suggest a semi-automatic hybrid approach covering several steps to build star schemas compliant to the data-source and decision-makers requirements simultaneously. The proposed approach uses a general semantic resource (i.e., WordNet) to overcome the semantic heterogeneity issues. Using such a resource alleviates the design time and cost because efforts for the development of a specific ontology are no longer necessary. Furthermore, being hybrid our approach leads to: i) define two classes of heuristics for extracting multidimensional components starting either from a relational data source and from user
requirements, and ii) match star schemas issued from user requirements and from the data source model; for this semi-automatic matching we need defining specific metrics to decide which schemas are similar (i.e., analyzing the same organization business process) and hence should be matched. The usage of metrics is a promising idea to optimize the matching process. Currently, we are finishing the architecture for a framework that supports the proposed approach; very soon we will start the development of a software prototype that supports the proposed DW design approach.

<table>
<thead>
<tr>
<th>Work reference</th>
<th>Automatic</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>Yes</td>
<td>Global ontology</td>
<td>Multidimensional elements</td>
</tr>
<tr>
<td>[4]</td>
<td>Semi</td>
<td>Ontology based DB + requirements</td>
<td>DW Ontology model</td>
</tr>
<tr>
<td>[10]</td>
<td>Semi</td>
<td>Ontology based DB + decisional requirements</td>
<td>Semantic MD model</td>
</tr>
<tr>
<td>[57]</td>
<td>Semi</td>
<td>Requirements analysis + sources analysis</td>
<td>Graph oriented MD model</td>
</tr>
</tbody>
</table>

**TABLE 2. HYBRID APPROACH USING SEMANTIC RESOURCE**

**REFERENCES**


ACM international workshop on Data warehousing and OLAP. ACM, 2005.


[38] Mazón, Jose-Norberto, Juan Trujillo, and Jens Lechtenbörger. "Reconciling requirement-driven data warehouses with data sources via multidimensional normal forms". Data & Knowledge Engineering 63.3 (2007): 725-751.


[51] Introduction to requirements elicitation.Website: www.cs.iastate.edu/~cs362/handouts/handout02.pdf


