Ontology-based Data Warehouse Development Process

Helena Vranešić  
Information Systems Engineering, Johann Wolfgang Goethe University  
Frankfurt am Main, Germany  
vranesic@wiwi.uni-frankfurt.de

Lidia Rovan  
Department of applied computing  
University of Zagreb Faculty of Electrical Engineering and Computing  
Zagreb, Croatia  
lidia.rovan@fer.hr

Abstract. The Data Warehouse projects confront different communities of practice to create one body of knowledge and help increase companies' competitive advantage. This paper gives an overview of communication problems on boundaries between participating communities. We have focused on the potential relation between the early creation of language communities of the involved communities and lowering data warehouse projects' development costs. We provide theoretical foundations for ontology-based data warehouse development scheme proposed in this paper.

Keywords: data warehouse, communities of practice, language community, ontology, boundary object

1. Introduction

Designing data warehouse (DWH) involves participants from different communities of practice. We distinguish two major groups of participants which confront each other in DWH projects: (1) OSPs (operative system professionals) with knowledge of the legacy and source systems, and (2) BEDFs (business experts in decision-making fields). OSPs are in charge of maintenance and further development of data warehouse source systems. BEDFs make informed decisions, using data warehouse reports, to determine course of actions the company needs to take in order to stay competitive in the environment. Connection between these two practices is made by introducing a third practice, DWH professionals who are responsible for the development of the DWH.

Communities of practice are characterized as shared histories of learning [1, 2], known to create discontinuities between the ones who participate in the community’s work and the ones who do not [1]. These discontinuities are also revealed in development of DWHs. Specifically, DWH professionals are confronted with communication gaps on boundaries towards two communities of practice [1]: BEDFs, who define business requirements, and OSPs in charge of data sources required for fulfilling those requirements. Our observation is that misunderstanding caused by communication gap results in not reaching the deadline, which leads to exceeding the budget and increasing maintenance cost. We assume that creating mutual understanding between DWH professionals and communities of practice involved in data warehousing is a successful road toward domain DWH ontology. Considering benefits use of ontology provides in solving technical DWH implementation issues, we believe that creating such ontology might be a precondition for a successful and cost-efficient DWH.

The structure of the remainder of the paper is as follows. In the next section, we give possible theoretical foundations relating the concepts of negotiation of meaning and language communities with two forms of boundary connections in frames of DWH projects. We propose our DWH Project Improvement Scheme. Following this, we provide an overview highlighting expected technical benefits of ontology-based DWH design approach. Finally, we summarize our findings and limitations based on this research in progress.
2. DWH Projects’ Communication Gap: Crucial Issues

2.1. Negotiation of Meaning

Wenger [1] uses the term *negotiation of meaning* as: “the process by which we experience the world and our engagement in it as meaningful”. Thus, in that process which involves an interaction of two constituent processes, namely *reification* and *participation*, our personal meaning of the world around us is located [1]. *Participation* is a complex process that combines all kinds of relations: doing, talking, thinking, feeling, and belonging [1]. Through the process of *reification*, aspects of human experience and practice are congealed into fixed forms and given the status of objects. However, products of reification are not simply concrete, material objects to be shared and easily understand by others. Rather, they are reifications of our perceived meanings. Hence, participation and reification form a duality, indicating how a union of object and its meaning is only possible in a discussion with the reification-maker [1].

Both processes can contribute to the discontinuity of boundaries (e.g., a specific jargon is understood only by participants of a community where it is used), but can also create continuities across boundaries. Thus, the process of reification produces objects that can move across boundaries [1]. Likewise, members of one community of practice can participate in multiple communities of practice at once and help to introduce those reified objects from one community of practice to the others, when such translation is needed.

2.2. Boundary Connections

According to several authors [1, 3], there are two forms of connections on the boundary between communities, vindicating roles of participation and reification: (1) *boundary objects* and (2) *brokering*. Star and Griesemer [4] define boundary objects as “both plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain a common identity across sites”. In different social contexts (i.e., communities of practice), they can have different meanings; however, their structure is common enough across contexts to make them recognizable by means of translation. In the context of DWH projects, the artifacts (e.g., shared documents, tools, business processes) exchanged between communities of practice [3] can potentially become boundary objects if they should belong to (at least) two different practices. When these artifacts as reifications of one community (exp. OSPs), want to be used by another (exp. BEDFs), they form a nexus of different perspectives (Figure 1.), which then needs to be coordinated [1].

![Figure 1. Creation of Language Communities between (OSPs & DWHs) and (BEDFs & DWHs)](image)

Brokering provides missing coordinating connection between communities of practice and includes activities by individuals that involve: facilitating transactions and the flow of knowledge between communities of practice [5]. In DWH projects, DWH professionals have brokering roles. Wenger [1], identified three types of brokering processes: the ability to *translate* elements of the “world view” of one community in terms of the perspective of another community, to *coordinate* different perspectives to a state where communities of practice are not working at cross-purposes and to *align* the enterprise of the communities of practice to a larger context of meaning. DWH professionals translate, coordinate and align different perspectives on both boundaries between OSPs and BEDFs and prevent communication gaps. To fill in gaps, DWH professionals (participative role) actively facilitate a negotiation of meaning process between all involved parties with help of boundary objects:

1. On one hand, gathering requirements in DWH projects means not only gathering created reports and documents, but also meeting with community’s representatives. Reports do represent and are related to some of the more tactical decision-making requirements [6], but they are purely reifications of them. Only through interaction with members of the BEDF community, where those objects were reified, can a full understanding of the requirements be reached. These objects then form the nexus where different perspectives of BEDFs and
DWH professionals through processes of participation. This suggests that communication between both practices is of paramount importance for success of DWH projects.

(2) On the other hand, in interaction with members of the OSP community, DWH professionals concentrate on the main process of DWH development: implementation of extraction, transformation and loading (ETL) of operational data with the goal of creating more appropriate data-sets matching the BEDFs’ requirements. The costs resulting from the ETL process have a major share in the overall DWH project cost [7]. Extraction of operational data implies the knowledge of: a) what operational data exactly needs to be extracted to meet the requirements and business demands, and b) which operative system this data needs to be extracted from. In this context, we suggest that communication between OSPs and DWH professionals is a profound prerequisite for aligning different perspectives of exchanged boundary objects (e.g., database schema, database tables, excel tables, text files etc.) in those two communities. Only in close cooperation with OSPs, DWH professionals can extract and interpret operational data so that it matches the requirements of the BEDFs.

2.3. Language Communities

As part of our research we propose that DWH professionals generate a mutual understanding of artifacts on the nexus of perspectives (i.e., boundary objects) through communication with members of both OSPs’ and BEDFs’ communities. In accordance with Ågerfalk and Eriksson [8], we too argue that traditional research on requirements analysis has focused too much on the syntactic and semantic, and too little on the pragmatic aspects of the requirements process. Language critique [9, 10], a branch of constructive philosophy known as the “Erlangen School” provides useful insights and backup for our understanding of DWH projects, and has been successfully applied to Information Systems (IS) research before [e.g., 11, 12-14].

Kamlah and Lorenzen [9] argue that language as a system of signs promotes mutual understanding as a "know-how" held in common, the possession of a "language community". A new term is introduced by explicit agreement between language users with respect to its usage and meaning [9]. This agreement leads to a relation of concept and term, and is shared by a language community as the knowledge of using this term. According to our understanding, boundary objects play a significant role in making language communities explicit: if members of communities of practice have a same concept in mind when confronted with a same artifact (boundary object), they belong to the same language community.

3. Ontology-based Design Approach

In order for DWH Professionals to acquire the necessary understanding of the involved communities of practice (OPS and BEDF), and thereby develop cost-efficient DWHs, we argue that DWHs - OSPs and DWHs - BEDFs language communities should be created in early stages of DWH development (Figure 2.)

We propose DWH Project Improvement Scheme (Figure 2.) as a conceptual development framework, incorporating negotiation mechanism for identifying potential problem areas as a first step in DWH development process.

Figure 2. DWH Project Improvement Scheme

During the process called negotiation of meaning, BEDFs and OSPs together with DWH professionals discuss the meaning of BEDFs’s set of business requirements and OSPs’s operational data sources supported by a kind of Analysis Panel. Artifacts used in process of negotiating of meaning (business documents, source system documentation etc.) are later gathered and stored in an Input Data Storage. Ambiguities in both, requirements and sources are considered as opportunities for repeated negotiation and creation of new understanding. Due to alternation of meaning, DWH professionals produce, vary and store a set of terms in the so-called Terms Storage. Each time participants meet (BEDFs and OSPs together with DWH) process of negotiation carries out and influences DWH professionals’ set of terms. When no novel versions are produced, term is stable and ready for usage. Language community is created when all necessary terms are stable.
For example, let’s consider the following scenario: Company X is in charge of data warehouse system development for an insurance company, in our case called A. During the first phase of requirements gathering, companies’ representatives, DWH professionals from company X, BEDFs and OSPs from company A, meet on regular basis. Most of the meeting sessions include discussions over company A’s legacy systems specifications as well as company X’s DWH solution proposition (both being afterwards stored in Input Data Storage). Those documents represent the foundation for negotiations of the meaning for concepts and terms involved (e.g., differences in the meaning of the terms “hospital” and “clinic”, that DWH professionals are unaware of, is cleared and both terms are afterwards stored in Term Storage). In case of existing medical ontology, negotiation of meaning process could have been accelerated by mapping: observed document-terms with predefined ontology-terms. Our proposed DWH improvement scheme could act as a framework that facilitates systematic investigation of hidden, weak and unexpected relationships between language terms from the perspective of visual exploration and analysis of discussed documents.

According to Rosenkranz and Holten [15], the language perspective of Kamlah’s and Lorenzen’s language critique approach can be used to create an ontology. Hence, pragmatic use of terms DWH professionals gathered in the creation process of language communities represents a good starting point for creation of DWH development process ontology. We believe, ontology could enforce further business requirements and operational data source collection as well as development of DWH solution and later it’s adaptive, perfective and corrective maintenance.

3.1. Ontologies

Ontology is a formal, explicit specification of a shared conceptualization. A conceptualization refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined [16]. Compared with other classification schemes, such as taxonomies, thesauri, or keywords, according to Huhns and Singh [17], ontologies allow more complete and precise domain models.

3.2. Creating ontology

There are several proposed methodologies and built tools for creating ontologies. Simpler and Tempich [18] give an overview of ontology engineering methodologies, while Nigam and Mehrotra [19] give an comparison of available ontology tools and emphasize on observed issues. Each of proposed methodologies works well for a particular scenario [19]. Imposing the pre-step of language community creation in DWH design, we believe we have successfully approached main issues in an iterative ontology development process: a) Identification of ontology domain and scope (What is the domain of DWH reports? Who are DWH users?) b) Retrieval of suitable existing ontologies c) Retrieval of necessary concepts and d) Definition of classes, properties and relationships.

Ontologies are intended to be shared and reused, moreover that approach is perceived to be beneficial [20]. Considering the number of publicly available pre-built ontologies and limited scope they usually cover, we believe finding a suitable ontology that would completely cover domain and scope of DWH is not feasible. Global DWH ontology should be built by reusing and merging existing ontologies, partially covering its domain. Simpler and Tempich [18, 21] emphasize the limitations of the techniques available so far, related to language translators and ontology merging. As response to prevailing ontologies problems, like inconsistency, incompleteness and redundancy of existing domain ontologies Janiesch [21] suggests two solutions: standardized engineering process for ontologies or resolution of domain differences already at a language level.

Existence of language communities is a natural precondition for ontology development process. Created language communities (DWHs-BEDFs and DWHs-OSPs) enable clear definition of an ontology domain and scope, retrieval of necessary concepts and definition of classes, properties and relationships between them.

The manual acquisition of ontologies still remains a tedious, cumbersome task resulting easily in a knowledge acquisition bottleneck [22]. Therefore, it is essential to propose appropriate automatization of ontology development process. Process starts in pre-design phase during the establishment of language communities. Terms determined during that phase form a subset of ontology concepts. Different process steps involve different team members. For example,
discovering necessary concepts and defining their structure and relationships involves all participating communities and usage of boundary objects while ontology engineering involves only DWH professionals.

One of the main advantages of ontology-based design is having syntactically correct and semantically consistent model. Reasoning over ontologies also provides retrieval of additional rules that were possibly not recognized during the design phase.

4. Additional value

Ontology includes machine-interpretable definitions of basic concepts in the domain and relations among them. Following DWH technical aspects could be easily accomplished with a help of ontologies: data integration (part of ETL) and interoperability.

According to our approach, in early stage of DWH design, global domain-specific ontology is developed. This ontology can be used to resolve semantic data integration problems in DWH projects. Data warehouse integrates and stores a copy of data from various sources that may differ in structure (e.g. relational databases, object databases, unstructured data, etc.). Even when sources have the same structure, data integration problems such as problem of synonyms and homonyms occurs [23]. When integrating data from different sources, the ability to distinguish among synonyms, homonyms and related terms is essential. Ontologies are used to ensure that two concepts, which might appear in different databases in different forms with different names, can be described as truly equivalent (i.e. they describe the same object). Research in the Semantic Web field resulted in numerous tools and methodologies for giving meaning to data [24, 25]. For easier integration, DWH data sources should be semantically enriched hence provided with well defined meaning using appropriate semantic integration technologies (e.g. D2RQ1 for relational database). In our case this means mapping source data to its matching meaning in predefined DWH ontology.

Interoperability is one of the most important features of modern applications. Semantic web technologies together with developed DWH global ontology enable proper annotation of resulting data (e.g. DWH report data). In this scenario semantic enabled tools could be used to perform additional functions on presented DWH data (e.g. Resource Description Framework attributes (RDFa) aware browsers). For example, exporting presented data as an input to other application is straightforward as all data could have assigned meaning using RDFa tags.

Introducing ontology to DWH design process was initialized primarily to define concise DWH model. Besides solving social problem of DWH design (misunderstanding between participants) ontology also provides great support to crucial technical problems: data integration and data interoperability.

5. Conclusion

Our prime goal was to enlighten communication problems emerging in DWH projects. We have shown how principles of creating a mutual understanding between all participants could lead to the successful DWH domain ontology development. In the process of the domain ontology creation DWH professionals have a unique and moderating role since they simultaneously belong to both communities of practice and to both language communities, which makes them aware of the different perspectives that OSPs and BEDFs may have. We expect that once created domain ontology could have an important role in DWH implementation process. During DWH project, developers are confronted with technical problems of data integration and interoperability.

We have indicated benefits ontology-based technologies could bring to that area. At this point one of the main limitations of our research is related to the versioning of continuously changing DWH domain ontologies.

By eliminating communication problems in an early project phase, DWH professionals obtain syntactically and semantically correct DWH model. Furthermore, taking into account additional technical improvements (introducing ontology to DWH design process) we proposed a possible approach to the development of a successful and cost-efficient DWH.

6. Acknowledgments

The presented work was jointly funded by German Federal Ministry of Education and Research, record no. 01FD0611 and by Croatian Ministry of Science, Education and Sports, record no. 036-0361983-201. The authors would

---

1 http://www4.wiwiss.fu-berlin.de/bizer/d2rq/
like to thank the funding agencies for their support.

7. References