Bridging the Gap between Data Warehouses and Organizations

Veronika Stefanov*

Women's Postgraduate College for Internet Technologies
Institute of Software Technology and Interactive Systems
Vienna University of Technology
stefanov@wit.tuwien.ac.at

Abstract. Data Warehouse (DWH) systems are used by decision makers for performance measurement and decision support. Currently the main focus of the DWH research field is not as much on the interaction of the DWH with the organization, its context and the way it supports the organization's strategic goals, as on database issues. The aim of my thesis is to emphasize and describe the relationship between the DWH and the organization with conceptual models, and to use this knowledge to support data interpretation with business metadata.

1 Problem Statement and Research Question

Data Warehouse (DWH) systems represent a single source of information to analyze the development and results of an organization[1]. Measures such as the number of transactions per customer or the increase of sales during a promotion are used to recognize warning signs and to decide on future investments with regard to the strategic goals of the organization.

Currently, the main focus of the DWH research field is on database issues, such as view maintenance, aggregation of data, indexing, data quality, or schema integration[2]. What has not yet been considered appropriately is the context of the DWH, its interaction with the organization and the way it supports the organization's strategic goals. The conceptual models in Data Warehousing are strongly data-orientated[3] and do not allow for formally describing DWH context. Models that describe the DWH from various viewpoints, including an outside view of the DWH system, its environment and expected usage, are missing. Moreover, eventhough the data in the DWH by its very nature has to be closely related to the concerns of the organization, current DWHs also lack sufficient business metadata that would inform users about the organizational context and implications of what they are analyzing[4].

This PhD proposal targets the relationship between the DWH and the organization with two interrelated research questions:

* This research has been funded by the Austrian Federal Ministry for Education, Science, and Culture, and the European Social Fund (ESF) under grant 31.963/46-VII/9/2002.

How can the relationship between the Data Warehouse and the structure, behavior, and goals of the organization...
(1) be formally described?
(2) support the interpretation of data?

Section 2 describes the research goals and the research field, followed by the expected results and their evaluation in Sect. 3, the contribution and beneficiaries of the expected results in Sect. 4, and a time plan and potential risks in Sect. 5. Section 6 describes the preliminary results achieved so far, followed by related work (Sect. 7), and a conclusion (Sect. 8).

2 Research Goals, Field, and Scope

I address the research questions stated in Sect. 1 with two goals:

1) Development of a Conceptual Modeling Language. Diagrams that show how the organization is related to the DWH will be developed, to make it possible to model how the organization interacts with the DWH, and how its structure and behavior are mirrored by the DWH (data) structure.
2) Creation of Business Metadata. Knowledge about the organization, captured in an enterprise model, will be linked to the DWH by means of model weaving [5] and used to gain business metadata. Business metadata describes the business context of the data, its purpose, relevance, and potential use[4].

These goals represent different ways of applying the same knowledge about the relationship between the DWH and the organization, and they achieve different contributions (see Sect. 4). Because this thesis applies modeling techniques to the DWH as the application area, it positions itself in a multidisciplinary research field between Model Engineering and Data Warehousing, as visualized in Fig. 1.

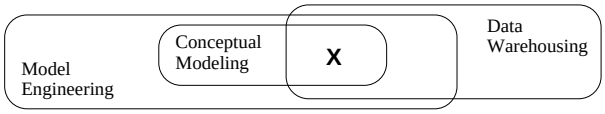

Fig. 1. Research Field of the PhD

The scope of this thesis is limited to the conceptual level and *the relationship between the DWH and the organization* only. It does not include DWH development projects (which have their own goals and also interact with the organization), technical details of data mapping and DWH design or methodology.

3 Methodology and Evaluation

The goals of the PhD will be achieved and the results evaluated as follows:

Development of a Conceptual Modeling Language. To reach the first goal, conceptual models to show the relationship between the DWH and the structure, behavior and goals of the organization will be developed. Models for five different aspects are planed. The models will be based on UML 2.0 and implemented as UML Profiles (preliminary results in Sect. 6.1).
Conceptual Models are difficult to evaluate. Related approaches in the area of DWH (see Sect. 7) are usually applied to examples and scenarios. Serrano et al. [6] attempt to empirically evaluate DWH data models with quantitative metrics. Wolff and Frank [7] propose a multi-perspective framework for evaluating conceptual models with regard to organizational change.
The preliminary results described in Sect. 6.1 were tested with example business processes. As soon as more mature models are available, I am planning to test them in a real-world setting at a bank, where a colleague has already expressed interest.

Creation of Business Metadata. To achieve the second goal, weaving models[5] will be developed to link conceptual models with the DWH data model. Through the weaving links, business metadata can be generated (for preliminary work, see Sect. 6.2.2). A prototype of a tool for creating weaving models and generating business metadata will be developed and tested on a real-world DWH.

4 Contributions and Beneficiaries

Conceptual models bring benefits during the earlier phases of the DWH lifecycle, such as requirements analysis and design, whereas business metadata supports the operational phase. Modeling how the organization interacts with the DWH, and how its structure, behavior and goals are mirrored in the DWH provides (1) *Increased Visibility* and (2) *Improved Communication*. This is useful during development of a DWH, leading to (3) *Facilitated Requirements Analysis*, (4) *Requirements-driven Design* and (5) *Streamlined DWH Evolution and Re-Engineering*. It also supports (6) *Documentation* and (7) *Maintenance*.

Business Metadata provides background information directly in the DWH, leading to (1) *Improved Data Interpretation* as well as (2) *Enhanced Usability and User Acceptance of Gathered Data*.

The beneficiaries of this thesis are therefore (a) all people involved in designing, building and maintaining a DWH (i.e. the architects and designers as well as the users). Their tasks are facilitated, and their project communication is improved by capturing volatile and implicit knowledge, and making it visible. And (b), during the operational phase of a DWH, users and maintainers benefit from improved interpretation through business metadata.

5 Time Plan and Risks

I plan to finish my PhD thesis by the end of 2007. This year (2006) is dedicated to developing additional conceptual models and the business metadata weaving models.

Among the risks of this PhD thesis are the interdisciplinary subject coupled with an unconsolidated understanding of the nature of Data Warehousing, which leads to a small immediate community, as well as the uncertain availability of suitable real-world examples.

6 Preliminary Results

This section gives an overview over the already completed parts of the thesis. Section 6.1 addresses research question 1 and presents a modeling approach for the relationship between DWHs and Business Processes. It is an excerpt of three papers that have already been published[8–10]. Concerning research question 2, Sect. 6.2 presents a weaving model based on an enterprise goal model.

6.1 Data Warehouses and Business Processes: A Conceptual Model

DWH information is accessed by business processes. Conceptual models can make the relationship between the DWH and the business processes visible. The *UML Profile for Business Intelligence (BI) Objects*[8] allows to show where and how a DWH is used by business processes, and which parts of the business processes depend on which parts of the DWH. We defined seven types of BI objects , representing the different types of data repositories, as well as the data models and the means of presentation of the data. Figure 2 shows an example process using the stereotypes "Fact" and "DWH". The BI objects are defined as stereotypes in a UML profile. The use of the stereotypes is guided by OCL constraints[11] provided with the profile, which can be automatically checked by many modeling tools.

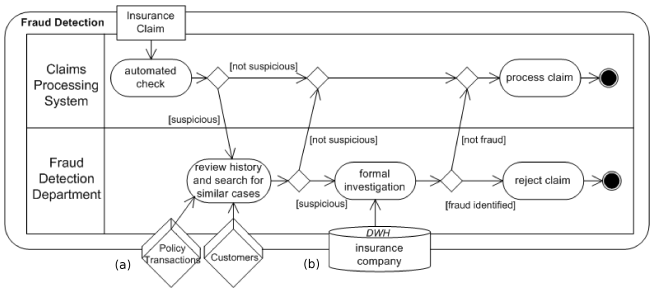

Fig. 2. Example fraud detection process, modeled as UML 2.0 activity diagram with BI Objects: Subprocesses access data from (a) two facts and (b) the whole DWH[8]

In [9], we investigated the relationship between DWHs and business processes from the viewpoint of performance measurement. DWHs provide Key Performance Indicators (KPIs), also called metrics or performance measures in other disciplines, that are accessed by business processes.

The *Performance Measurement Perspective* is an extension to the Event-Driven Process Chain (EPC)[12]. It provides model elements for KPIs and other performance measurement capabilities of a DWH environment.

Finally, [10] offers a broader look at the relationship between DWHs and business processes, as it also takes active, real-time DWHs into account. We presented a two-fold approach that adds two perspectives to the EPC. In addition to the *Traditional BI Perspective*, which contains modeling elements for a classic DWH environment, the *Active BI Perspective* allows to model how an active DWH influences the control flow of a business process.

Regarding related work, many business process models include features to show data access, but they do not take the special characteristics of DWH data into account.

6.2 Business Metadata concerning Enterprise Goals

In order to provide business metadata in the DWH, the context of the DWH, i.e. the structure, behavior and goals of the organization, has to be modeled in an Enterprise model. This model is then weaved with the data model of the DWH, to create links for metadata.

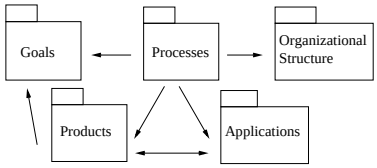

Fig. 3. A basic enterprise model

6.2.1 Enterprise Model Enterprise models are used to formally represent the structure, behavior and goals of an enterprise organization. They are usually organized into separate aspects[13].For example, an organization chart can be used to describe the departments, groups and roles that exist within the organization, and a business process model to describe the structure of business processes. Figure 3 shows the outline of a basic enterprise model, organized into five packages. The business metadata to be created is aimed at covering all areas of the Enterprise model. “Enterprise model” is used here in a much wider sense than commonly in Databases, where the term often denotes enterprise *data* models.

6.2.2 A Weaving Model between Enterprise Goals and the Data Warehouse Model The first approach to create business metadata for DWHs exploits the relationship between decision support and enterprise goals. What is good or bad performance, and which decisions should be taken based on the data, depends on the goals to be reached. Enterprise goals concern market share, inventory levels or customer satisfaction and can be seen as an abstraction of business structure and behavior, as they form the basis for decisions and the way a company does business. They govern the design of business processes and the way the organization behaves.

We introduce weaving links[5] between a multidimensional data metamodel (a simplified form of [14]) and an enterprise goal metamodel as shown in Fig. 4. The links of the weaving model can be used to gain business metadata for the DWH, such as in the example in Tab. 1.

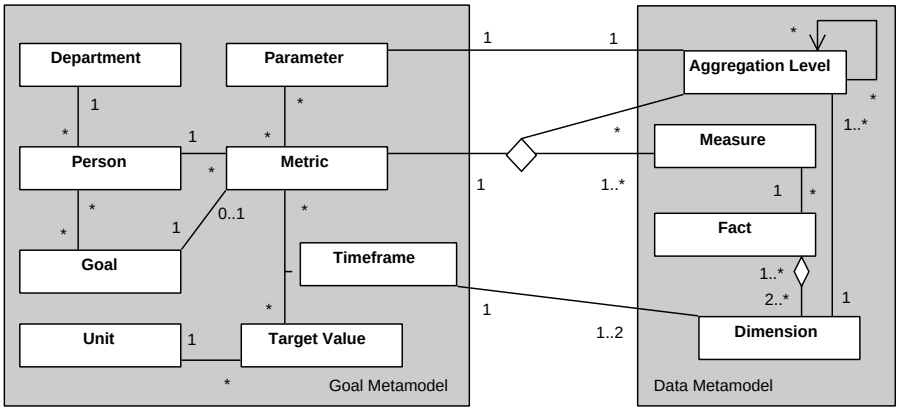

Fig. 4. Three weaving links between enterprise goals and multidimensional metamodels

The central link in Fig. 4, connecting the *Metric* of a goal with a *Measure* from the DWH (and optionally with an *Aggregation Level*), can be explained as follows: In the goal model, a metric measures the degree of fulfillment of a goal (e.g. goal "reduce inventory cost" was reached to 80%. The metric with its target value and timeframe is related to the corresponding DWH measure, i.e. "inventory cost" of the *Fact* "Inventory", which supplies the actual values. When accessing the measure, the weaving link allows to access all the information recorded in the enterprise goal model, e.g. who the metric is reported to or which goal it corresponds to. The upper link relates aggregation levels to the *Parameters* of a metric, whereas the third link connects the *Timeframe* of a metric's *Target Value* to the *Dimensions* containing temporal values in the DWH.

There are many approaches to support DWH design with goal modeling [15, 16]. But, the goals analyzed in these cases are either goals of the DWH itself (e.g. data quality, usefullness, availability) or goals of the DWH project (e.g. timeliness), but not goals of the enterprise organization.

Metric name:	Reduction of inventory cost
Target value + unit:	100 Euro
Responsible + contact info:	Ms. Smith, ext. 51564, ...
Reported to (person/dept.) + contact info:	Ms. Baker, ext. 51324, ...
Goal supported by this metric:	reduce inventory cost
Optional: Conflicting or supporting goals:	conflict: "provide on-time delivery"

Table 1. Example business metadata for the measure "inventory cost"

7 Related Work

The approaches described in this PhD proposal are in line with requirements-driven DWH design. Approaches to DWH design generally fall into two main categories[16, 17]. *Data-driven* (also *supply-driven* or *bottom up*) approaches focus on the data sources that are available. The main question is how this data can be extracted and transformed into a multidimensional data model. *Requirements-driven* (also *demand-driven* or *top down*) approaches on the other hand instead use the user requirements and enterprise goals as a starting point[18], and leave the identification of data sources to a later phase.

Conceptual modeling in the area of Data Warehousing has largely focussed on database related areas, namely the data model and schema transformations. The main data model in Data Warehousing is the multidimensional model, also called star schema[19]. It is meant to provide intuitive and high performance data analysis[1]. There are many approaches to modeling the multidimensional data structures of DWHs (for comparisons, see [20]). The structure of the data model of a DWH is relevant to this work only in terms of relating and connecting it to other models, in order to enrich the DWH with business metadata.

Linking DWH business metadata with technical metadata to provide a better context for decision support was first suggested in [4]. Several business metadata categories and a number of desirable characteristics are defined. The business metadata is described with UML classes and associations and linked directly to technical metadata within the same model. The approach only covers metadata and does not include separate conceptual models of the business context.

8 Conclusion

DWH systems are used by decision makers for performance measurement and decision support. Since the main focus of the research field is on database issues, most effort has been put on improving on how the DWH works, and the question how it is used has mostly been neglected so far.

In this thesis, I propose to use conceptual models for describing the relationship between the DWH and the structure, behavior, and goals of the organization, to increase the visibility of this relationship and to improve communication by capturing this knowledge. Moreover, business metadata can be added to the DWH that informs users about the context and background of the data, in order to improve data interpretation.

References

1. Kimball, R., Reeves, L., Thornthwaite, W., Ross, M., Thornwaite, W.: The Data Warehouse Lifecycle Toolkit. John Wiley & Sons, Inc. (1998)
2. Vassiliadis, P.: Gulliver in the land of data warehousing: practical experiences and observations of a researcher. In: Proceedings DMDW'00, CEUR-WS.org (2000)
3. Rizzi, S.: Conceptual Modeling and Evolution in DWs. Perspectives Workshop: Data Warehousing at the Crossroads, Dagstuhl, August 1-8 (2004)
4. Sarda, N.L.: Structuring Business Metadata in Data Warehouse Systems for Effective Business Support. CoRR (2001)
5. del Fabro, M.D., Bézivin, J., Jouault, F., Breton, E., Gueltas, G.: AMW: A Generic Model Weaver. In: Proceedings IDM'05. (2005)
6. Serrano, M., Calero, C., Trujillo, J., Luján-Mora, S., Piattini, M.: Empirical Validation of Metrics for Models of Data Warehouses. In: Proceedings CAiSE'04, Springer-Verlag Heidelberg (2004) 506–520
7. Wolff, F., Frank, U.: A Multi-Perspective Framework for Evaluating Conceptual Models in Organisational Change. In: Proceedings ECIS'05. (2005)
8. Stefanov, V., List, B., Korherr, B.: Extending UML 2 Activity Diagrams with Business Intelligence Objects. In: Proceedings DaWaK'05. LNCS 3589, Springer (2005) 53–63
9. Stefanov, V., List, B.: A Performance Measurement Perspective for Event-Driven Process Chains. In: Proceedings DEXA 2005, IEEE (2005) 967–971
10. Stefanov, V., List, B., Schiefer, J.: Bridging the Gap between Data Warehouses and Business Processes: A Business Intelligence Perspective for Event-Driven Process Chains. In: Proceedings EDOC '05, IEEE Computer Society (2005) 3–14
11. Object Management Group, Inc.: UML 2.0 Object Constraint Language (OCL) Specification. http://www.omg.org/cgi-bin/apps/doc?ptc/05-06-06.pdf (2005)
12. Keller, G., Nüttgens, M., Scheer, A.W.: Semantische Prozeßmodellierung auf der Grundlage "Ereignisgesteuerter Prozeßketten (EPK)". Veröffentlichungen des Instituts für Wirtschaftsinformatik (89) (1992)
13. Whitman, L., Ramachandran, K., Ketkar, V.: A taxonomy of a living model of the enterprise. In: WSC '01, IEEE Computer Society (2001) 848–855
14. Luján-Mora, S., Trujillo, J., Song, I.Y.: Extending the UML for Multidimensional Modeling. In: Proceedings UML '02, Springer-Verlag (2002) 290–304
15. Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P.: Fundamentals of Data Warehouses. Second edn. Springer-Verlag New York, Inc. (2001)
16. Giorgini, P., Rizzi, S., Garzetti, M.: Goal-oriented requirement analysis for data warehouse design. In: Proceedings DOLAP 2005, ACM (2005) 47–56
17. Winter, R., Strauch, B.: A Method for Demand-Driven Information Requirements Analysis in Data Warehousing Projects. In: Proceedings HICSS 03, IEEE (2003)
18. Prakash, N., Gosain, A.: Requirements Driven Data Warehouse Development. In: CAiSE Short Paper Proceedings, Springer (2003)
19. Chaudhuri, S., Dayal, U.: An Overview of Data Warehousing and OLAP Technology. SIGMOD Rec. **26**(1) (1997) 65–74
20. Abelló, A., Samos, J., Saltor, F.: YAM^2 (Yet Another Multidimensional Model): An Extension of UML. In: Proceedings IDEAS '02, IEEE (2002) 172–181