A Common Meta-Model for Data Analysis based on DSM

Yvette Teiken   Stefan Flöring

OFFIS - R&D Division Health
Escherweg 2
26121 Oldenburg, Germany
teiken@offis.de   floering@offis.de

Abstract

Realization and usage of advanced decision support systems are cost intensive. They require expert users during data collection and analysis task. Implementing these systems is time consuming and thus costly. This leads to the problem that SMEs (Small and Medium-sized Enterprises) often cannot afford these systems. In this paper we describe our aim of creating a DSM (Domain Specific Modeling) based top-down approach to generate advanced decision support systems. This approach is based on a family of DSLs (Domain Specific Language) that share a common meta-model. With this we aim to establish a faster and more affordable process for data analysis which better fits for SMEs.

Keywords DSM, DSL, Data Analysis, Visual Analytics, Decision Support Systems

1. Introduction

In our research group Data Management and Analysis we are dealing with the subjects collection, storage and analysis of complex multidimensional data. For this purpose we developed MUSTANG (Multidimensional Statistical Data Analysis Engine), a platform to implement specialized analytical information systems based on a data warehouse (Koch et al. 2003). It is, for example, used in the epidemiological cancer registry of Lower Saxony (Rohde and Meister 2004). The MUSTANG platform supports the collection and analysis of multidimensional data to supply information and decision support.

In this paper we describe a new top-down approach to data analysis based on DSM to extend the MUSTANG platform.

2. Brief overview of our research activities.

Available platforms for data analysis are often not optimized for specific domains or analysis approaches. Instead, they supply general solutions which require the work of a domain expert who establishes the analysis environment. Besides, the expert tries to find a visualization method which will fit the given analysis purpose best.

From our perspective this approach has two major shortcomings. First of all it requires that a large amount of visualization methods are integrated into the analysis platform. While this may technically not be challenging, it requires the data analysis expert to have an overview over a possibly large amount of visualization methods, of which many might not fit his needs. Another shortcoming results out of the data model which is used by the analysis platform. Multi-dimensional data models reclining the OLAP data model are typically used. This approach is of practical use because of the wide spread usage of systems which can provide data in this kind of model.

However, the allocation of indicators, derived from the multidimensional data model, to axis or other characteristics of a given visualization-method is not trivial. It requires deep semantic knowledge of the data. Data analysis therefore is a task limited to those users who exhibit this knowledge.

Both challenges lead to the conclusion that only a specialized data analyst with deep knowledge of both, the analysis domain and the technical characteristics of the visualization, can accomplish analysis tasks in an analysis platform like this. Reports generated by the analysis platform are then used as decision guidance. Therefore, the whole analysis process typically involves multiple users: data analysis experts and decision-makers. This procedure fits the needs of larger enterprises where data analysis and decision making are generally shared between different compartments due to the corporate hierarchy.

SMEs often do not distinguish between management-level and expert-level. Additional man power for analysis tasks is considered to be too cost intensive. This occurs in particular in healthcare environments which are under considerable cost-pressure. Nonetheless, there are vast amounts of data recorded in many healthcare applications which are a valuable source for data analysis. A streamlined and more cost efficient warehousing process which requires less knowledge by the analyst might help to establish warehousing in SMEs.

2.1 Model Driven MUSTANG

More and more organizations are using decision support systems like performance management (Friedrich 2007). Some of these approaches, e.g. Analytical Performance Management (Koch 2008), use a data warehouse (DWH) for data storage. A DWH is a central data storage from different sources. The data is used for data analysis for supporting decisions in organizations. The classical DWH process is a data driven approach. Analysis is based on existing data (Mucksch 2006). For supporting management decisions a demand driven approach is requested (Martin and Nußdorfer 2007). In a demand-driven process a decision maker formulates questions which shall be answered by the decision support system.

For integrating this kind of process in DWH a top-down approach must be realized. Therefore, we are creating a demand-driven top-down approach in MUSTANG. To accomplish this goal we want to define multiple DSLs as done in (Warmer and Kleppe 2006). A schematic diagram for our process is given in figure 1. This shows how an information demand in the context of epidemiology will be modeled and conditioned in our approach. In our approach a domain expert will be able to model relevant issues, his information demand, in a Measure DSL. These issues are expressed in key figures so the domain of this DSL is modeling key figures. The DSL will be used to model and as far as possible generate OLAP cube models. An OLAP cube is used to store a set of measures with dimensional context (Codd et al. 1993) and is used

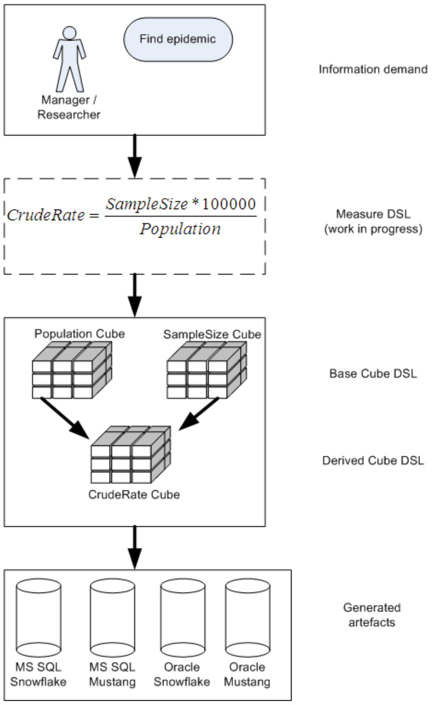

Figure 1. Process for data management in Mustang

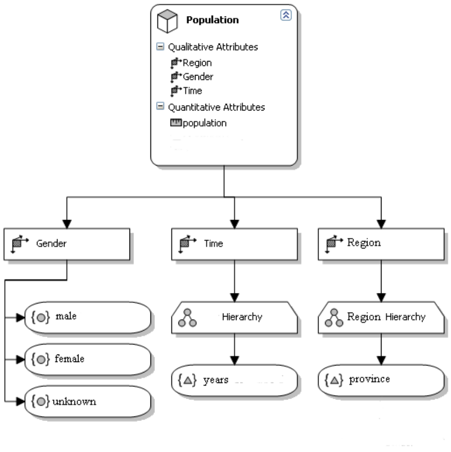

Figure 2. Example instance of our cube DSL

for data management. These models are multidimensional views on key figures with different levels of aggregation and are based on OLAP cubes. The cube itself is described by a cube DSL. In regard to the cube model the cube DSL is a more technical view on the domain and will be used as supplement of the Measure DSL. For better reuse and to benefit from a well-known language the DSL is based on ADAPT (Bulos 1996). Measures and dimensions are described graphically in the DSL. An example for a cube modeled in our cube DSL is given in figure 2. Based on this model we can generate a multidimensional data schema represented in SQL scripts. By means of a DSM approach we are able to support various databases and multidimensional schemata and applications for data integration based on the cube model. We can generate different kinds of multidimensional schemata like snowflake-, star- and MUSTANG-Schemata based on a cube-model. Supported databases are currently MS SQL and Oracle. This variety of data models is conditioned by the heterogeneous environments of our different research partners.

In future research we want to add software management tools like SAMA as described in (Koch and Teiken 2008). SAMA is a DSL based approach of monitoring strategy maps (Teiken 2008). We also want to add other DSLs to enable full generation of decision support systems and DWH. This will include a DSM based approach for security, integration and data quality.

2.2 Visual MUSTANG

Model driven MUSTANG is a first step for creating a streamlined, cost efficient data analysis solution. However, an approach for visualization is missing. We argued that domain experts with deep knowledge about both, visualization characteristics and data semantics, are required to choose appropriate visualization method. Our idea to bear this challenge is to use a visualization DSL to describe certain aspects of visualization methods, such as available dimensions or available operators on the visualized data. In (Hanrahan 2006) a language for this purpose, VizQl (visualization query language) is introduced. VizQl allows declarative queries for visualizations in a similar manner as SQL for data. We use this description to generate task or domain specific visualization applications which fit best for a certain analysis purpose.

We identified two steps in this process. One step which is technically motivated and domain independent and another step which is knowledge based and requires a domain expert. The first step is to identify technical characteristics of visualization models like the minimum and maximum number of dimensions and the available operators a certain visualization method provides. This is a technical task, which will result in a description language for the characteristics of visualization methods. In addition to this we want to be able to describe the explorative operators supported by a visualization method for a certain characteristic. In a map diagram, for example, a zoom-in operation for the visualization can intuitionally be used as drill-down operation.

The second step is more sophisticated. Here the knowledge of a domain expert about which kind of visualization is a good (or the best) choice for certain parameters of a given data model is needed. For example for one analysis task values with time dimension may be best visualized as an animation of diagrams. An example for this is the GapMinder [1]. Another task might require displaying time as index on the x-axis of a coordinate system.

Those two descriptions form a visualization DSL, which allows to describe the visualization methods best fitting to certain data model or analysis task. A similar DSL for data models is necessary to create a software development process which will allow automatic generation of visualization applications for certain data models. A similar approach is used in (Bull 2006) where model driven visualization is used to rapidly prototype new visualizations.

3. A Common Meta-Model for data analysis

In our joint work we noticed that many of these research questions may be solved with the help of a DSM based approach based on a common meta-model as recommend in (Hessellund et al. 2007). Our different approaches will be realized by using different DSLs. We call those related DSLs a family of DSLs. We aim for our family of DSLs to share a common meta-model. The reason why we use different DSLs rather than different view points is variousness of our reception radius. We address our DSLs to different kinds of domain experts like a manager or security expert. In case of cube and dimension modeling it would be appropriate to use view points.

[1] http://www.gapminder.org

From this common model we expect synergies in the development of a streamline process for all phases in data analysis, and thus a more cost effective realization of projects which will allow SMEs to take advantage of data analysis.

For every DSL a meta model to express its abstract syntax is essential (Völter and Stahl 2006). For the integration of our DSLs we decided to use a common meta-model. A simplified version of the MUSTANG DSL meta-model is shown in figure 3. The cube DSL is connected to the Measure DSL via measure. A cube can store a single measure in different granularities. This is called a base measure. A measure can also be a combination of other measures which is called a derived measure. Each cube has a number of operations and dimensions with hierarchies. An operation describes potential OLAP operations for a cube. The dimension has a type that describes special properties, e.g. if it's a geographical dimension. An example of a simplified instance of our meta-model is given in figure 4. It shows how the information demand "find epidemic" in figure 1 with help of crude rate is expressed. One application of these properties can be the type of visualization of the dimension.

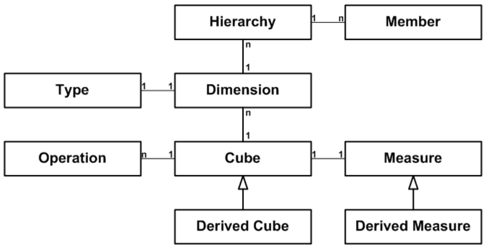

Figure 3. Part of MUSTANG Meta-Model

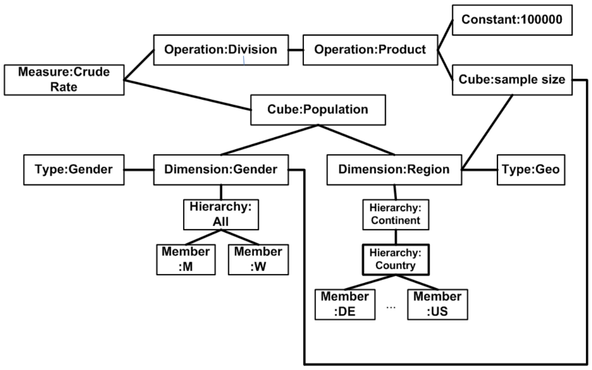

Figure 4. Instance of our meta-model

A meta-model for the visualization DSL needs to include the following two items: presentable characteristics of the visualization method (dimensions, hierarchies, members) with value margins and operators supported by the visualization. In our research we discovered that the meta-model which is used for the cube DSL can be used for the visualization DSL as well if the annotations mentioned are added. Therefore, we decided to use a common meta-model for both languages.

Figure 5 shows how visualizations described by the visualization DSL are matched to an instance of the MUSTANG meta-model. At first the characteristics of the data model are detected. After this step the amount of available visualizations is limited to those which support the characteristics of the data model. In the next step annotations of the data model which are derived from expert knowledge are evaluated. In the described process of generating a top-down decision support system the model incorporates knowledge by expert users. This knowledge is available to the matching process. This matching process results in the best matching visualization for the application generation. If more than one visualization is applicable to a certain task, there are two choices: Either an expert user can manually influence the generation process and choose one of these visualizations or an application with the support of multiple visualizations will be generated.

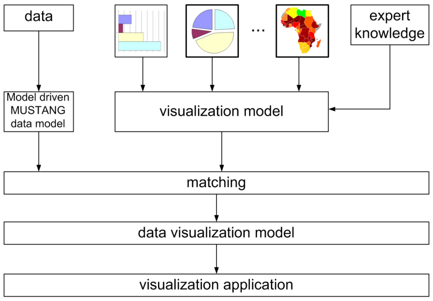

Figure 5. Visualization Matching Process

4. Conclusion

Using the DSM based approach a decision support system with appropriate visualization can be realized cost efficient. With our family of DSLs we can generate suitable applications based on models described by domain experts with a minimum of manual programming.

Another advantage of this approach is a possible higher user satisfaction. When modeling measures and cubes with the specific DSLs we generate domain knowledge. This domain knowledge can be used to choose not only a technically fitting but a semantically appropriate visualization.

On the other hand due to the DSM approach it is easier to create and integrate additional visualizations into MUSTANG. The characteristics of a visualization method may be described and applications can be adapted without implementation work afterwards. Using the common meta-model the integration of data can be fully defined within the model, reducing the developers work for data integration. These two aspects open the perspective for a DSM based roundtrip-engineering (Antkiewicz and Czarnecki 2006) where exploration and visualization of data might imply adjustments to the model which with help of our family of DSLs may automatically be propagated to the model.

References

Michal Antkiewicz and Krzysztof Czarnecki. Framework-specific modeling languages with round-trip engineering. In Oscar Nierstrasz, Jon Whittle, David Harel, and Gianna Reggio, editors, *MoDELS*, volume 4199 of *Lecture Notes in Computer Science*, pages 692–706. Springer, 2006. ISBN 3-540-45772-0.

R. Ian Bull. Integrating dynamic views using model driven development. In *CASCON '06: Proceedings of the 2006 conference of the Center for Advanced Studies on Collaborative research*, page 17, New York, NY, USA, 2006. ACM. doi: http://doi.acm.org/10.1145/1188966.1188989.

Dan Bulos. Olap database design: A new dimension. *Database Programming&Design*, Vol. 9(6), 1996.

Edgar F. Codd, Sharon B. Codd, and Clynch T. Salley. Providing olap to user-analysts : An it mandate. White paper, E.F. Codd Associates, 1993.

Dirk Friedrich. Einfach soll es sein - bei hoher datenqualität. *IS - Informationsplattform für Business Applications*, 11:30–35, 2007. (in german).

Pat Hanrahan. Vizql: a language for query, analysis and visualization. In *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 721–721, New York, NY, USA, 2006. ACM. ISBN 1-59593-434-0. doi: http://doi.acm.org/10.1145/1142473.1142560.

Anders Hessellund, Krzysztof Czarnecki, and Andrzej Wasowski. Guided development with multiple domain-specific languages. In Gregor Engels, Bill Opdyke, Douglas C. Schmidt, and Frank Weil, editors, *MoDELS*, volume 4735 of *Lecture Notes in Computer Science*, pages 46–60. Springer, 2007. ISBN 978-3-540-75208-0.

Sascha Koch. *Analytisches Performance Management*. PhD thesis, Universität Oldenburg, 2008. (in german).

Sascha Koch and Yvette Teiken. Semi-automatische überwachung von zielsystemen. In Martin Bichler, Thomas Hess, Helmut Krcmar, Ulrike Lechner, Florian Matthes, Arnold Picot, Benjamin Speitkamp, and Petra Wolf, editors, *Multikonferenz Wirtschaftsinformatik*. GITO-Verlag, Berlin, 2008. ISBN 978-3-940019-34-9. (in german).

Sascha Koch, Jürgen Meister, and Martin Rohde. MUSTANG – A Framework for Statistical Analyses of Multidimensional Data in Public Health. In *Proceedings of the 17th International Conference Informatics for Environmental Protection*, pages 635–642, Cottbus, September 2003. (in german).

Wolfgang Martin and Richard Nußdorfer. Cpm – corporate performance management, kompendium:analytische services in einer soa, teil 1: Herstellerunabhängige beschreibung und referenzarchitektur. White Paper, August 2007. (in german).

Harry Mucksch. *Analytische Informationssysteme. Business Intelligence-Technologien und -Anwendungen: Business Intelligence-Technologien Und -Anwendungen*, chapter Das Data Warehouse als Datenbasis analytischer Informationssysteme, pages 129–142. Springer, Berlin, 2006. (in german).

Martin Rohde and Jürgen Meister. *Data-Warehouse-Systeme. Architektur, Entwicklung, Anwendung*, chapter Data Warehousing in der Gesundheitsberichterstattung, pages 484–498. dpunkt Verlag, 2004. (in german).

Yvette Teiken. *Semi-automatische Überwachung Annotierter Strategy Maps*. 2008. ISBN 3899636724. (in german).

Markus Völter and Thomas Stahl. *Model-Driven Software Development*. Wiley & Sons, 2006.

J. B. Warmer and A. G. Kleppe. Building a flexible software factory using partial domain specific models. In *Sixth OOPSLA Workshop on Domain-Specific Modeling (DSM'06), Portland, Oregon, USA*, pages 15–22, Jyvaskyla, October 2006. University of Jyvaskyla. ISBN 951-39-2631-1.