

## Tools for Data Warehouse Quality

M. Gebhardt, M. Jarke, M. A. Jeusfeld, C. Quix, S. Sklorz  
 Informatik V, RWTH Aachen, Ahornstr. 55, 52074 Aachen, Germany  
 {gebhardt,jarke,jeusfeld,quix,sklorz}@informatik.rwth-aachen.de

### Abstract

In this demonstration, we show three interrelated tools intended to improve different aspects of the quality of data warehouse solutions. Firstly, the deductive object manager ConceptBase is intended to enrich the semantics of data warehouse solutions by including an explicit enterprise-centered concept of quality. The positive impact of precise multidimensional data models on the client interface is demonstrated by CoDecide, an Internet-based toolkit for the flexible visualization of multiple, interrelated data cubes. Finally, MIDAS is a hybrid data mining system which analyses multi-dimensional data to further enrich the semantics of the meta database, using a combination of neural network techniques, fuzzy logic, and machine learning.

### 1. Introduction

Quality factors such as accessibility and timeliness, believability and understandability, design and usage flexibility play a crucial role in the success of data warehousing. The European ESPRIT Long Term Research Project DWQ (Foundations of Data Warehouse Quality [9]) attempts to address these issues in a systematic manner, and to link design options for specific data warehouse components and policies to an overall architecture and quality model [8].

The DWQ project is developing a number of prototypical tools to illustrate the improvement potential of our approach. The tools described in this short paper focus firstly on the aspects of metadata management, and secondly on improving client-side interaction with data warehouses supporting a rich multidimensional data model. Aspects of data refreshment and source integration are only marginally addressed, because they are mainly covered by other partners in the project.

In section 2, we describe how ConceptBase, a metadata management system supporting a deductive object model, can be used to handle a semantically oriented metamodel of data warehouses and to support explicit quality management via this metamodel. In section 3, we present CoDecide, a visually oriented multi-dimensional data model by which geographically distributed teams of users can rapidly construct and change views over networks of data cubes. Finally, section 4 presents a more automated way of data analysis which also supports fur-

ther enrichment of metadata semantics: the MIDAS system combines neural network techniques for unsupervised clustering with a fuzzy learning component and a novel visual analysis interface. In the conclusions, we sketch the linkage to other aspects of data warehouse quality.

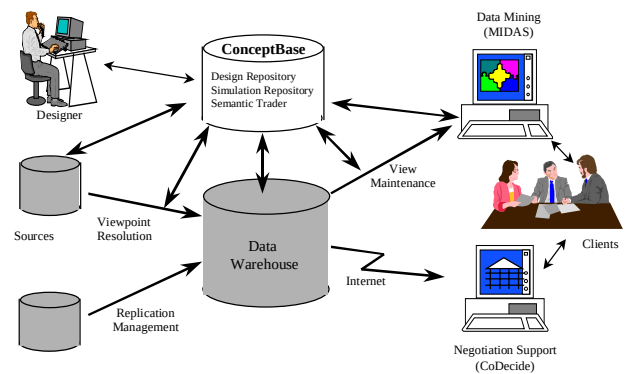


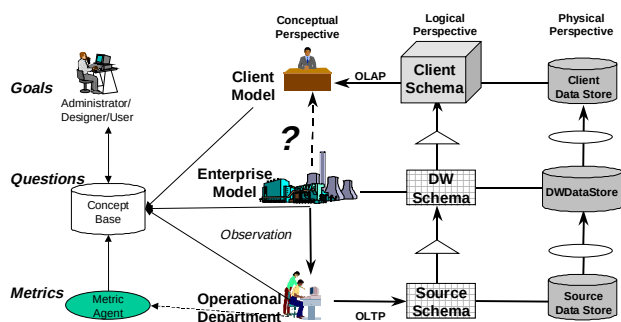
Figure 1: Role of demonstration in DW setting

### 2. Metadata Management with ConceptBase

ConceptBase is a meta database manager intended for conceptual modeling and co-ordination in design environments. It integrates techniques from deductive and object-oriented databases in the logical framework of the data model Telos [7]. The meta-modeling ability of Telos allows designers to represent heterogeneous modeling languages like ER diagrams or UML. Objects described in one modeling language can be linked to objects in some other modeling language. Rules and constraints expressed as logical formulas can encode the axioms of the respective language. The meta class hierarchies of ConceptBase have unlimited extensibility. Meta classes, classes and instances can co-exist in the same object base and queries can be used to examine the classes stored in ConceptBase.

Many aspects of data warehouses have been studied in database research, including materialization and maintenance of views, integration of legacy sources, and modeling of multidimensional data. However, the current data warehouse meta models cannot express the large number of quality factors of data warehouses. The consequence is, that there is no systematic understanding of the interplay between quality factors and design options in data warehousing.

In the DWQ Project, we have developed an architectural and quality management framework, that is implemented in ConceptBase. This framework extends the standard data warehouse architectures by modeling also enterprise aspects. We have adapted the Goal-Question-Metric (GQM) approach [14] from software quality management in order to link these techniques to our conceptual framework of a data warehouse. The idea of GQM is that quality goals can usually not be assessed directly, but their meaning is circumscribed by questions that need to be answered when evaluating the quality. Such questions again can usually not be answered directly but rely on metrics applied to either the product or process in question.



### Figure 2: Managing Data Warehouse Quality with GQM

ConceptBase is used as a metadata repository for information about the architecture of the data warehouse as well as a model to store quality parameters of each data warehouse component and process [8]. The query language of ConceptBase can be used to analyze a data warehouse architecture and its quality, e.g. to find out weaknesses and errors in the design of a data warehouse.

The implemented solution uses a similar approach as GQM to bridge the gap between quality goal hierarchies on the one hand, and very detailed metrics and reasoning techniques on the other. The bridge is defined through quality measurements as materialized views over the data warehouse architecture and through queries over these quality measurements. The measurements are stored in the ConceptBase repository by external metric agents, e.g. a tool for measuring the response time or a reasoner for checking the consistency and minimality of the data warehouse schemata. The queries of ConceptBase are used to evaluate the stored measurements and give an evidence for the fulfillment of certain quality goals.

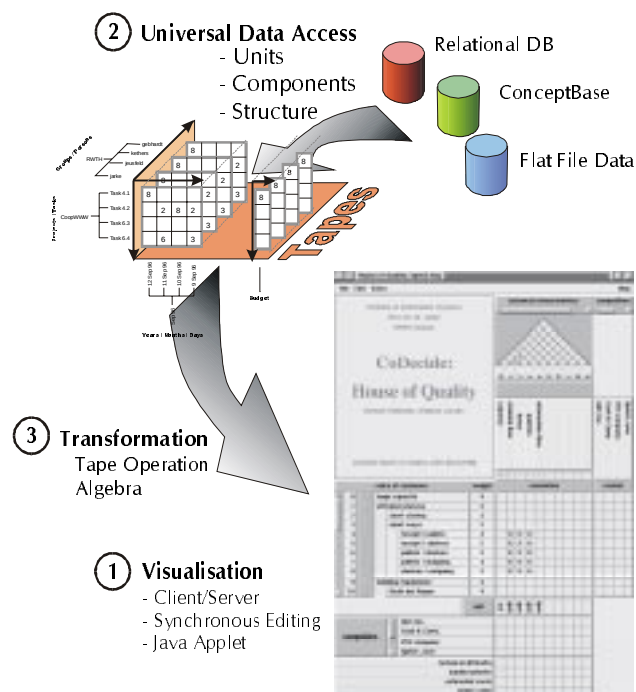
Our implementation strategy gives more technical support than usual GQM implementations and allows the reuse of existing technologies for assessing and optimizing the quality factors of a data warehouse. The current work focuses on the stabilization of the quality model, the integration of external metric agents with ConceptBase and the examination of quality factors in a data warehouse.

### 3. Analysing Interlinked Data Cubes with CoDecide

The basic idea of OLAP is to support decision making by presenting the relevant information based on up-to-date data retrieved from various data sources. The multi-dimensional approach allows to focus quickly on relevant information cubes e.g. by slice and drill down operations. But one problem remains: It is difficult to visualize the connection between two or more such information cubes.

CoDecide is an experimental user interface toolkit using a novel visualization technique for interlinked, multidimensional data which handles this problem.

In CoDecide the multi-dimensional data is broken up into inherently 2-dimensional building blocks called tapes. Any analytical perspective could then be constructed by interactively composing and transforming the tapes to CoDecide worksheets (cf. (1) in figure 3). In contrast to the pivot table approach used, e.g. in Excel [22], we do not construct a single matrix from the involved dimensions. Instead, we arrange multiple matrix segments within tapes, thus creating a family of inter-linked views on the problem. These views can be looked at (e.g. scrolling, drill-down/roll-up) and manipulated (e.g. adding information) together. Moreover, they can be distributed across workstations with different access rights to the overall structure and different degrees of synchronization, thus enabling a wide variety of cooperative support options.



**Figure 3: CoDecide OLAP Architecture**

A second difference is that hierarchy operations such as drill-down and scroll-up are not separated out in a design interface but directly embedded in the matrix

interface, taking advantage of the opening/ closing option familiar from the Macintosh interface. Thus, CoDecide can be thought of as a user interface equivalent of the Data Cube operator, except that multiple related views on a cube or even multiple data cubes ("galaxy model") are supported.

A local area network version of CoDecide has been operational since 1994 [4]. It was used in a number of design support applications with moderate database sizes, including factory layout planning [5] and business process analysis [10]. This version was implemented using Tcl and C on an X11 platform. It maintained its own multi-dimensional data cache, and allowed real-time synchronous collaboration ('what you see is what I see') for up to about five simultaneous negotiators on a LAN.

The demonstration shows a new Java-based version of CoDecide which is used to assist in cooperative decision making on the WWW. Two applications developed with CoDecide (time management in projects, and program committee management) have been integrated with the BSCW workspace system [2] as part of another European project, CoopWWW [1].

Interfaces for downloading from external data sources had initially to be built on a one-by-one basis. CoDecide has now a heterogeneous database interface [6], [18] to interactively access data in relational, object-oriented and flat file databases (cf. (2) in figure 3). To fill the gap between the data model used to store the data persistent in a database system, and the data model used to answer a specific analytical questions, the database interface is complemented by a tape algebra [12] (cf. (3) in figure 3).

#### 4. Generating Fuzzy Metadata with MIDAS

The diversity of data mining methods and the large number of different mining tasks exert a conspicuous influence on the development of data mining tools. Information about several dozens of such tools, research prototypes as well as commercial products, is available on the KDNuggets Web site [15]. The trend is that tools are designed either to cope with only a single mining task (single-strategy tools) or to provide an arsenal of different mining techniques in one integrated environment (multi-strategy tools). Single-strategy tools benefit from the possibility to adapt mining techniques to specific applications and are therefore often used to support mining in dedicated application areas. Multi-strategy tools contain several mining techniques in order to fulfill the requirements of a variety of applications. However, even these tools cannot be prepared for all potentially occurring mining problems, considering the rapidly increasing number of most different data mining applications.

Some developers of multi-strategy tools react to this general problem by proposing extensible tool architectures allowing data mining methods to be integrated to the system whenever required. This presupposes an ex-

tensive software engineering, but it should not dominate data mining activities. What else can be done? Every discipline involved in data mining research provides techniques with specific computational properties, such as recognition of patterns, explanation of decisions or handling of numerical data. The strength of a single technique is often the weakness of another one, and vice versa. More attention could be given to suitable combinations.

The multi-strategy tool MIDAS is such a profitable combination of mining techniques. The central idea is the combination of neural networks with decision tree induction methods. Both techniques are ubiquitous in data mining. A brief outline of their interaction illustrating the substantial components of MIDAS is given in figure 4.

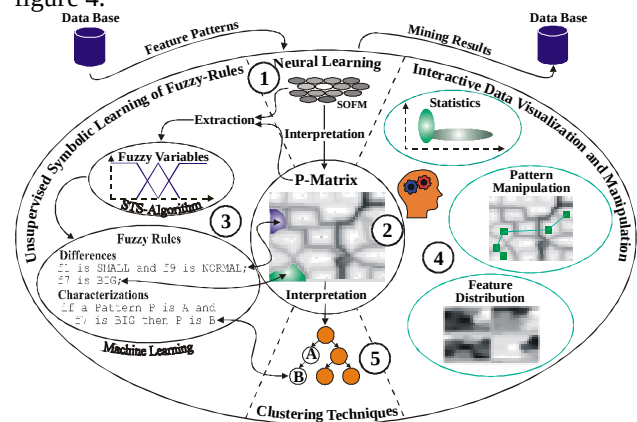


Figure 4: MIDAS Overview

Input for MIDAS are feature patterns from the space  $R^m$  which are used to train an unsupervised neural network named self-organizing feature map (cf. (1) in figure 4). This neural network proposed by Kohonen [11] maps the input patterns to a two dimensional grid of units so that the relative positions of the mapped patterns in the grid show their similarity in the input space.

MIDAS directly uses this essential quality to discover an a-priori unknown number of clusters. This is done with the help of a graphical interpretation method called P-Matrix [16]. The P-Matrix enables the user to identify interactively different regions in the grid that are separated by dark borders in the graphical display of P (cf. (2) in figure 4). Each region corresponds to a cluster of the input patterns. The system supports this interpretation with algorithms that interpret P in order to build up a hierarchy of clusters automatically (cf. (5) in figure 4).

After clusters have been discovered, MIDAS can be used to learn cluster descriptions (cf. (3) in figure 4). A number of fuzzy-terms is generated for every feature of the input patterns. The neural network's ability of generalization is central in this step called *signal-to-symbol* (STS) which transforms the numerical data into a fuzzy-logical representation [17]. The idea is to create fuzzy-rules consisting of generated fuzzy-terms in order to describe the discovered clusters. In principle, such de-

scriptions can be learned with machine learning methods which are able to operate on fuzzy terms and on a set of positive and negative examples. Up to now, we have integrated a decision tree induction method named fuzzy-ID3 [20].

Besides the automatic rule generation, MIDAS offers several statistical and graphical modules to support an interactive data analysis (cf. (4) in figure 4). The system is implemented in Java and can be connected to a data base via a JDBC/ODBC bridge. A comprehensive description of the system and its architecture can be found in [21],[13].

MIDAS has been used successfully in two commercial projects, namely the analysis of retail data and of city population descriptive data. The first project was carried out in cooperation with a German retail chain and the second one was a project with the municipal administration of a big German city.

## 5. Conclusion

In this short paper, we have described three aspects of tool support for improved design and usage quality in data warehousing. The main emphasis of ongoing work in the DWQ project is twofold: to further enrich the semantics of metadata based on advanced reasoning techniques over formal models of source integration and multidimensional data views [3], and to develop qualitative as well as quantitative methods and toolkits for the improved design of data warehouses at the logical and physical level, taking the conceptual knowledge into account [19].

**Acknowledgments.** This work was supported in part by the commission of the European communities under contracts DWQ and CoopWWW, and by the DFG through Aachen's focused Ph.D. programme on Computer Science and Engineering. We thank the partners in these projects for many valuable discussions.

## References

- [1] W. Appelt, "CoopWWW – Interoperable Tools for Cooperation Support on the World-Wide Web", Final Report, Project TE 2003 Telematics Applications Programme
- [2] R. Bentley et al., "Basic Support for Cooperative Work on the World Wide Web", *Intl. Journal on Human-Computer Studies*, spring 1997
- [3] D. Calvanese, G. DeGiacomo, M. Lenzerini, "Conjunctive query containment in Description Logics with n-ary relations", *International Workshop on Description Logics*, Paris, 1997.
- [4] M. Gebhardt, "Coherent Design by Coupling of Views", *Diploma Thesis* (in German), RWTH Aachen, 1994
- [5] M. Gebhardt, S. Jacobs, "Conflict management in design" *8th Intl. Symposium on Quality Function Deployment*. Detroit, MN, 1996.
- [6] M. Gebhardt, M. Jarke, S. Jacobs, "A Toolkit for Negotiation Support Interfaces to Multi-Dimensional Data" *Proc. ACM SIGMOD Conf.*, Tucson, AZ, 1997
- [7] M. Jarke, R. Gellersdörfer, M.A. Jeusfeld, M. Staudt, S. Eherer, "ConceptBase – a deductive objectbase for meta data management", *Journal of Intelligent Information Systems*, 4, 2, 167-192, 1995.
- [8] M. Jarke, M.A. Jeusfeld, C. Quix, P. Vassiliadis, "Architecture and Quality in Data Warehouses", *Proc. of the 10th Conference on Advanced Information Systems Engineering (CAiSE '98)*, Pisa, Italy, June 1998.
- [9] M. Jarke, Y. Vassiliou, "Foundations of data warehouse quality – a review of the DWQ project", *Proc. 2nd Intl. Conf. Information Quality (IQ-97)*, Cambridge, Mass., 1997.
- [10] M. Jarke, S. Gebhardt, S. Jacobs, H.W. Nissen, "Conflict analysis across multiple viewpoints: formalization and visualization" *Proc. 29th Hawaii Intl. Conf. System Sciences*, Wailea, Hw, 1996, vol. III, 199-210.
- [11] T. Kohonen, "Self-organized formation of topologically correct feature maps", *J. Biological Cybernetics* 43, 59-69, 1982
- [12] R. Mesbahipour, "Eine Algebra für die Restrukturierung des Bandmodells", *Diploma Thesis* (in German), RWTH Aachen, 1997
- [13] M. Mücke, "Linguistische Interpretation selbstorganisierender Merkmalskarten unter Berücksichtigung von Unschärfe", *Diploma thesis* (in German), RWTH Aachen, 1997
- [14] M. Oivo, V. Basili, "Representing software engineering models: the TAME goal-oriented approach", *IEEE Trans. Software Engineering.*, 18, 10, 1992
- [15] G. Piatetsky-Shapiro and M. Beddows, "KDNuggets Directory: Data Mining and Knowledge Discovery Resources", <http://www.kdnuggets.com/>
- [16] S. Sklorz, "A Method for Data Analysis based on Self-Organizing Feature Maps", *Proc. of the World Automation Congress (WAC'96) -- Soft Computing with Industrial Applications*, Vol. 5, TSI Press, Albuquerque, pp. 611-616, May 1996
- [17] S. Sklorz and M. Mücke, "A Hybrid Approach for Medical Data Analysis", *Proc. of the 5th European Congress on Intelligent Techniques and Soft Computing (EUFIT)*, Vol. 2, pp. 1162-1166, Aachen/Germany, September 1997
- [18] R. Stössel, "Iterative Projektion externer Daten in das Bandmodell", *Diploma Thesis* (in German), RWTH Aachen, 1997
- [19] D. Theodoratos, T. Sellis, "Data Warehouse Configuration", *Proc. 23th VLDB Conference*, Athens, 1997.
- [20] R. Weber and H.-J. Zimmermann, "Automatische Akquisition von unscharfem Expertenwissen", *KI Journal* (in German), pp. 20-26, 1991
- [21] G. Wirtz, "Interaktive Datenanalyse auf Basis selbstorganisierender Merkmalskarten in Datenbanken", *Diploma thesis* (in German), RWTH Aachen, 1997
- [22] Microsoft Excel User Manual. Microsoft Inc., 1994.