PROCESS REFERENCE MODEL FOR DATA WAREHOUSE DEVELOPMENT

A consensus-oriented approach

Ralf Knackstedt, Karsten Klose, Björn Niehaves, Jörg Becker
European Research Center for Information Systems, University of Muenster, Leonardo-Campus 3, Münster, Germany

Keywords: Data Warehouse development, Management Information Systems, Process Reference Model, Theoretical Foundation, Epistemology

Abstract: IS literature provides a variety of Data Warehouse development methodologies focussing on technical issues, for instance the automatical generation of Data Warehouse or OLAP schemata from conceptual graphical models or the materialization of views. On the other hand, we can observe a growing influence of conceptual modelling in the move of general IS development which is specifically addressing early phase design issues. Here, conceptual modelling solves communicational problems which emerge when for instance IT personnel and business personnel work together, mostly having distinct educational and professional backgrounds as well as using distinct domain languages. Thus, the aim of this paper is to provide the foundation of a Data Warehouse development methodology in form of a process reference model which is based on a conceptual modelling approach.

1 INTRODUCTION

A broad variety of methods and procedural or phase models aims at supporting the Data Warehouse development process (Poe 1996, Hammergren 1996, Hackney 1997, Devlin 1997, Inmon & Imhoff & Sousa 1998, Golfarelli & Rizzi 1999). Nevertheless, the lack of epistemological funding of research methods and methodologies is apparent and extensively discussed in the IS discipline (cp. for example Hirschheim & Klein & Lyytinen 1995, Keen 1980, Mingers 2001). Therefore, the aim of this paper is do provide a process reference model, a methodology, for Data Warehouse development under special consideration of its theoretical-epistemological fundaments.

We now specifically analyzed the consensus-oriented approach on conceptual modeling (Niehaves & Klose & Knackstedt & Becker, 2005) which is characterized by an interpretivist position, which is mainly colored by the critical linguistic approach of (Kamlah & Lorenzen 1973). The information models developed contain formalized linguistic statements to be tested for validity in combination with additional (empirical) research methods. This is done through members of a linguistic community in order to obtain consensus. Therefore, elements of the semantic theory of truth (Tarski 1944) and the consensus theory of truth are considered and used.

In Section 2 we introduce the consensus oriented approach. In Section 3 we specify certain elements of the consensus-oriented approach and operationalize them in the form of a research process model. By this means, we are able to elucidate the consensus-oriented approach to a Data Warehouse development methodology in particular. We conclude with a summary and an outlook in Section 4.

2 THE CONSENSUS ORIENTED APPROACH

The consensus oriented approach as a underlying paradigm of our reference process model of data warehouse development is related with the assumption that there is a real world existing independently from human speech and thinking processes. Thus, we assume the ontological realism.

The approach aims to create a linguistic community. Linguistic communities can be created through the (re)construction of an ortho-language. First parts of the language can be formed by the alignment of individual (real world) objects to nominators. In the context of IS-development

Knackstedt R., Klose K., Niehaves B. and Becker J. (2005).
PROCESS REFERENCE MODEL FOR DATA WAREHOUSE DEVELOPMENT - A consensus-oriented approach.
In *Proceedings of the Seventh International Conference on Enterprise Information Systems*, pages 499-505
DOI: 10.5220/0002534104990505

important nominators are terms such as ‘customer Meier’, ‘product 4711’ etc. Based on nominators, predicators (in our context e. g. ‘customer’ and ‘product’) are introduced in order to expose and communicate similarities of individual objects. Language has in this case immense impact on a subject’s perceptional processes. It defines the very basic perception and differentiation system. Shared language means shared conceptualizations about the real world among the members of a particular linguistic community.

In the move of the consensus-oriented approach to Data Warehouse development, two distinct languages come into play: particular conceptual modelling languages as well as natural languages. Following Tarski (Tarski 1944) the creation of the linguistic community takes place on two levels. On the first level (here named T* object language) conceptual model statements are expressed. For instance, using Entity Relationship Models (ERM) members of the linguistic community have to agree upon the term ‘entity type’; in the case of Event Driven Process Chains (EPC) they have to agree upon the term ‘event’. Moreover, a distinction between a) the language of model instances and b) the language of the modeling method and technique has to be made. On the second level (here named T* meta language) members of the community have to agree on a language which facilitates them to debate about the truth and nontruth of the statements represented in a model (including for instance German or English). In the next step, the meta language T* is used to discuss the modeling system which is formulated on the first level using the T* object language until a consensus of a group of experts is achieved. Afterwards, the results can be evaluated within the scope of the interpersonal verification (Kamlah & Lorenzen 1973, Kamlah & Lorenzen 1996). The formalized linguistic statements contained in a conceptual model are logically decomposed (deduction) until they are accessible as elemental statements for purposes of truth verification. This takes place by means of a group of experts who obtain a consensus. The main instruments are observation, experiments, interviewing and the interpretation of texts (Kamlah & Lorenzen 1996). The validity of statements in the model can be confirmed, for example, in the case of business specific models, with a single case. In case of a pattern or reference model, however, the generalized abstraction of different individual verifications (induction) is necessary. Here, research methods such as field experiments, surveys, case studies or action research can be applied. Based on these results a revision of the conceptual model is required.

Furthermore, the consensus oriented approach results in the following epistemological positions: Both empirical statements (Kamlah & Lorenzen 1996) and a priori statements can be made, which may form the basis of conceptual models. Conceptual modeling therefore derives its results via theoretical reflection of the model contents, as well as from the implementation of the model in information systems and through observation.

Conceptual models are one form of artefacts of a formalized language and can contain both empirical and a priori knowledge. Both inductive and deductive conclusions can be accessed firstly in the context of the model creation and secondly in the context of truth verification.

3 DATA WAREHOUSE DEVELOPMENT METHOD

The consensus-oriented information modelling approach provides a suitable foundation for a methodological framework for the specification of Data Warehouse systems. In terms of a process reference model, types of tasks and procedure recommendations can be identified. Thereby, the creation of a linguistic community requires several steps which have to be performed:

On the T* object language level a selection, modification or development of an appropriate modelling language is necessary. Moreover, it has to be ensured that all project members have a common understanding of the model constructs used. The language-critique approach requires an explicit and consistent introduction of terms. At the outset, basic terms are introduced. Their definitions are based on familiar terms which are shared in the understanding of all project members. Normally, such generally known terms can be taken from natural language. Based on those terms, other terms are reconstructed stepwise (explicitly introduced) unless it can be assumed that their meaning is naturally known. Thus, the terms received from the normalised language are based on each other. But the derivation of a term A must not use terms B which in turn requires an introduction of term A. Hence, it is necessary to ensure that terms are not reused in different meanings.

For the multi-dimensional specification of Data Warehouse requirements, a broad variety of modelling techniques exists. The documentation of multi-dimensional modelling techniques is often not totally conforming to the demand of the language-critique approach. For example, the introduction of the modelling technique ADAPT of Bulos (Bulos

1996) (as many more do as well) emphasises a detailed explanation of notation aspects in the form of model examples and symbol lists. However, a step-by-step introduction of the modelling technique language constructs and their corresponding documentation in the form of language-oriented meta models is increasingly used and more and more applied in the domain of Data Warehousing (for example in the form of Entity-Relationship-Models and additional textual explanations for the intended semantics of the model elements).

An application of the language-critique approach which can easily be transferred to the modelling of Entity Relationship Models has been developed by Wedekind (Wedekind, 1992) for example. Among other things, he proposes the usage of the construct operators subsumption, subordination, and composition.

Using construct operators, core terms of Data Warehousing can systematically be introduced. This is presented by Holten by means of a modeling technique for the specification of management views in information warehouse projects (Holten 2003), for example. Referring to his modelling technique the following modelling constructs can be identified as highly relevant in the context of Data Warehousing.

According to (Riebel 1979) Dimension Objects are defined as all entities which can be related to a decision in a business process (such as products or customers). With respect to the analysing purpose, Dimension Objects are assigned to Dimensions, which arrange objects hierarchically (e.g. for the analysis of product groups). Hierarchies can be divided in several Hierarchy levels. Dimensions that comprise identical Dimension Objects as leaf elements are combined to Dimension Groupings. Dimension Scopes represent the selection of several Dimension Objects from a Dimension. A Dimension Scope Combination can be regarded as a navigation space of Dimension Objects that can be analysed by the operations aggregation and disaggregation with respect to the hierarchies of the combined Dimension Scopes. Ratios define important aspects of Dimension Objects such as invoice and payment amount, gross margin, profitability, etc. and can be organised in Ratio Systems, which define selected relationships between Ratios in a mathematical or business-logical sense. As business information can not be expressed exclusively based on Ratios or Dimension Objects, they are combined to Information Objects. They describe the amount of data (as combinations of Ratios and Dimension Objects which is called Fact), a manager should analyse by Dimensions and Ratio Systems.

A comparison of core terms used in the modelling technique by Holten with terms used in other modelling techniques (cp. Figure 1) emphasises the necessity of the construction of a language community (cp. Holten 2003, Holten & Dreiling & Schmid 2002 [MetaMIS], Sapia & Blaschka & Höfling & Dinter 1998, Hahn & Sapia & Blaschka 2000 [ME/RM], Bulos 1996 [ADAPT], Golfarelli & Maio & Rizzi 1998, Golfarelli & Rizzi 1999 [DFM]). On the one hand, the comparison underlines that identical modelling constructs are named in different ways (synonyms). On the other hand, different modelling constructs are allocated with idem terms (homonyms). Furthermore, several modelling constructs remain generally unconsidered or dissimilar considered (for example in combination or in a multiple application of several constructs) in the different modelling techniques.

MetaMIS	ME/RM	ADAPT	DFM
Dimension Object	-	Dimension Member	-
Dimension	-	Hierarchy	-
Hierarchy Leve	Dimension Level Entity Typ	Level	Dimension Attribut
Dimension Scope	-	Scope	-
Dimension Grouping	-	Aggregatior Dimension	Hierarchy
Dimension Scope Combinatior	-	-	-
Ratio	Attribut	Dimension Member	Fact Attribute (Measure)
Ratio System	-	Measure Dimension	-
Information Object	Fact Relationship Typ	Hypercube	Fact
-	Attribut	Dimension Attribut	Non Dimension Attribut

Figure 1: Comparison of Data Warehouse Modelling Techniques.

In case of a missing general understanding of terms used on instance level (for example the product lines ‘food’ and ‘non-food’), an introduction of these terms (in particular dimensions and ratios or ratio systems) by means of the language-critique approach is necessary as well. Synonyms and homonyms especially occur with respect to ratios in different enterprise departments (for example turnover with or without taxes). For the construction of language communities, glossaries and rules for the determination of ratios (developed in consensus) should be an integral element of a conceptual modelling technique.

Moreover, the construction of a linguistic community comprises an agreement on a language which makes it possible to discuss about the new developed or modified modelling language and its model artefacts. Normally, a selection of a natural language such as English, Spanish or German is sufficient. Against the background of consensus-oriented modelling, this language is called T* meta language. In the context of Data Warehouse systems specification, using T* meta language mainly aims to achieve a consensus between project members about necessary information needs.

In the context of consensus building, consensus-oriented modelling requires that project members speak the same language. Moreover, they have to be '*rational*' and '*competent*' in the project domain (cp in detail Kamlah & Lorenzen 1996).

In the context of Data Warehouse development, especially a consensus about information needs is required. Literature debates of information need analysis are often based on a model including several overlapping circles (cp. Figure 2(a)). Following Szyperski (Szyperski 1980), the first circle represents the amount of information which is

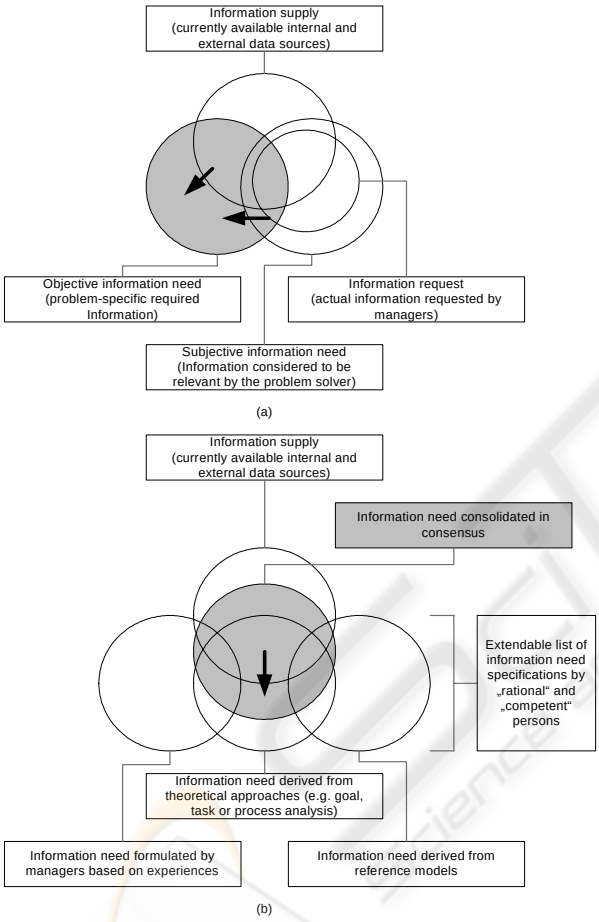

Figure 2: Information need analysis

actually available in an organization. The second circle represents the amount of information '*objectively*' needed for decision making or task performing. The third circle comprises the amount of information a user considers '*subjectively*' to be relevant for his/her task. Within the third circle, another circle exists which represents the amount of information that is explicitly demanded and articulated by a user. The model implicates that information represented in the first circle (information demanded by managers) and third circle (the available information) should be modified with respect to the "objective" information need (cp. the two arrows in Figure 2(a)).

In the context of consensus-oriented modelling, a discussion and representation of an '*objective*' information need is inappropriate, since in Data Warehouse projects exclusively information needs articulated and identified by project members are relevant. Following our presented epistemological position, this information need has to be interpreted as '*subjective*'. Thereby, a consensus about the "subjective" information need is required. On the one hand, different perceptions are based on dissimilar individual experiences of project members. On the other hand, they are particularly based on the methodical foundation of the formulated information needs. In this context, amongst others, the following three approaches can be distinguished (cp. Figure 2(b)):

- *End user involvement*: The participation of end users in the Data Warehouse development process is indispensable. Their involvement ensures the acceptance of the Data Warehouse system through the consideration of individual preferences and habits (for example the understanding of ratios). Manager often the information overload problem as they often demand the total amount of data available which is referred to a certain topic (Ackoff 1967). The differentiation between identified and articulated information needs made by Szysperski indicates that not only methodical conditions for an analysis of management information needs (for example in the form of observations, interviews, and surveys) have to be established. Furthermore, suitable cultural conditions are required which facilitate and motivate Data Warehouse users to express their information needs.
- *Methods for specifications of information needs*: The development of methods for a theoretically funded identification and specification of information needs has been a major issue in information systems research in the last decades. Several information requirements engineering methods and approaches (especially in the MIS domain) have been developed and evaluated (Martin 1983, Munro & Davis 1977, Sethi & Teng 1988, Rockart 1979). Nevertheless, the problem of information requirements engineering is considered to have deficiencies in theory.
- *Use of reference models*: Reference models are increasingly applied in the requirements

specification phase of Data Warehouse Projects. Reference models provide useful means to reduce the effort of information modelling, because they can be used as a starting point for the construction of project-specific models. Thus, reference models provide best (or common) practice solutions for information modelling projects. By means of reference models, opinions and experiences of external experts concerning the design of Data Warehouse systems are additionally and indirectly involved. The efficiency and effectivity of reference model applications can be increased by the use of configurative reference models.

Accounting for different advantages and disadvantages of the presented approaches, we propose a multi-methodological procedure. By this means, the acceptance of the Data Warehouse is ensured through the involvement of end users. Moreover, we overcome restrictions of an information need which is exclusively formulated by managers with an extended consideration of methods for the specification of information needs. Finally, reference models extend the (indirect) participation of additional experts.

Following the consensus oriented approach, a verification of the consensus consolidated is needed. For an inter-personal verification in the context of Data Warehouse projects, several artefacts (distinguishable in granularity and implementation orientation) can be taken into account. One possibility is to decompose requirements specification models in single statements. Afterwards, these statements are verified (for example relationships within dimensions of multi-dimensional models or the consistence of ratio formulations).

Another possibility is to implement a complete Data Warehouse system on basis of requirements specification models. This realisation may result in the fact that there is no longer a consensus about certain parts of the requirements specification models. Instead of a complete implementation of the Data Warehouse system, usually prototypes are developed and realised for verification purposes.

Based on the verification, it may be necessary to modify the information modelling results. These modifications may affect specific information models but also selected modelling methods (for example due to the fact that additional representations are required). Therefore, the T* object language is used again. For the discussion of necessary modifications, the T* meta language is applied.

4 SUMMARY AND OUTLOOK

We summarise our results concerning the development of Data Warehouse systems based on

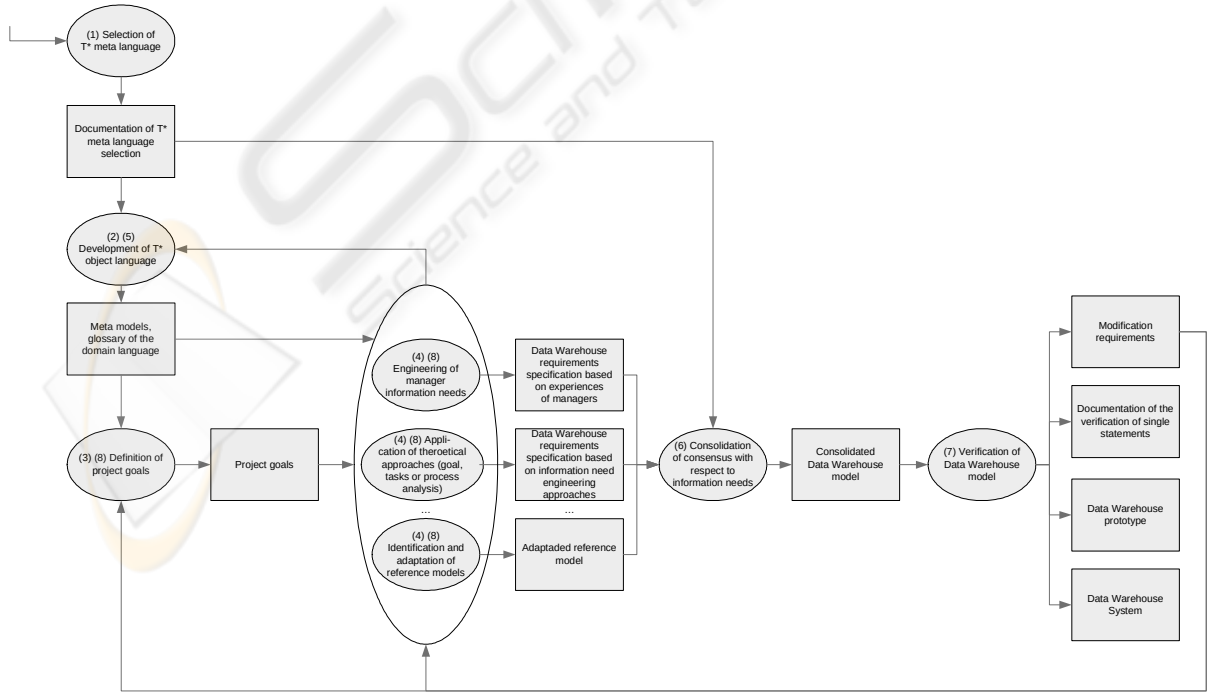

Figure 3: Process reference model for Data Warehouse development.

the consenus-oriented approach in terms of a process reference model in Figure 3. Ovals represent types of tasks. Rectangles symbolise types of documents which are assigned to tasks. Documents can be outputs which are created by tasks or represent inputs which are used for the accomplishment of tasks. Thereby, types of tasks may comprise other tasks. Numbers assigned to tasks illustrate the (reading) order of the process reference model. However, several tasks are cross-sectional since they are passed through more than once.

The project goal needs to be defined in advance to the requirements specification. It facilitates the coordination of parallel information need engineering. Moreover, it particularly describes the context of management tasks that should be supported by the Data Warehouse system. The project goal definition itself is a model which is represented in the T* object language which requires the development of a language community as well.

Our framework emphasises the phase of conceptual modelling of Data Warehouse projects. Logical and physical aspects are only addressed in the context of the interpersonal verification. Thus, our approach has to be combined with other works, which stress logical and physical aspects of Data Warehouse development.

In comparison with other existing Data Warehousing procedure model approaches, the presented framework uses the consensus-oriented approach of conceptual modelling as a specific theoretical foundation. Instead of practical argumentation or mathematical deductions, our approach is based on the language-critique philosophical work of Kamlah and Lorenzen. Their work is used as a basis, because it comprehensively addresses the communication problems between Data Warehouse project members. Furthermore, our approach emphasises the explication of its underlying epistemological assumptions, which is associated with the definition of the consensus-oriented modelling approach (cp. Niehaves et al., 2005; Niehaves, 2004).

REFERENCES

Ackoff, R. L., 14 (1967), S. B-147 bis B-156., 1967. *Management Misinformation Systems*. Management Science, 14 (1), B 147-156.

Bulos, D., 1996. *A New Dimension. OLAP Database Design*. Database Programming & Design, 9 (6), 33-37.

Devlin, B., 1997. *Data Warehouse. From Architecture to Implementation*. Reading, UK et al.

Golfarelli, M., Maio, D. and Rizzi, S., 1998. *The Dimensional Fact Model – A Conceptual Model for Data Warehouse*. International Journal of Cooperative Information Systems, 7 (2-3), 215-246.

Golfarelli, M. and Rizzi, S., 1999. *Designing the Data Warehouse: Key steps and crucial issues*. Journal of Computer Science and Information Management, 2 (3).

Hackney, D., 1997. *Understanding and Implementing Successful Data Marts*. Reading, UK et al.

Hahn, K., Sapia, C. and Blaschka, M., 2000. *Automatically Generating OLAP Schemata from Conceptual Graphical Models*. In Proceedings of the ACM Third International Workshop on Data Warehousing and OLAP (DOLAP 2000), Washington D. C., USA, 10. November 2000.

Hammergren, T., 1996. *Data Warehousing. Building the Corporate Knowledge Base*. London, UK et al.

Hirschheim, R., Klein, H. and Lyytinen, K., 1995. *Information Systems Development and Data Modeling: Conceptual and Philosophical Foundations.* Cambridge University Press. Cambridge/MA.

Holten, R., 2003. *Specification of Management Views in Information Warehouse Projects*. Information Systems, 28 (7), 709-751.

Holten, R., Dreiling, A. and Schmid, B., 2002. *Management Report Engineering – A Swiss Re Business Case*. In From Data Warehouse to Corporate Knowledge Center (E. Maur and R. Winter Ed.), 421-437, Springer, Heidelberg, Germany et al.

Inmon, W. H., Imhoff, C. and Sousa, R., 1998. *Corporate Information Factory*. New York/NY, U.S.A. et al.

Kamlah, W. and Lorenzen, P., 1973. *Logical Propaedeutic*. Lanham/MD.

Kamlah, W. and Lorenzen, P., 1996. *Logische Propädeutik. Vorschule des vernünftigen Redens.* 3 Edition. Stuttgart, Weimar.

Keen, P. G. W., 1980. *MIS Research: Reference Disciplines and a Cumulative Tradition*. In Proceedings of the First International Conference on Information SystemsEd.), 9-18, Philadelphia/PA.

Martin, E. W., 1983. *Information Needs of Top MIS Managers*. MIS Quarterly, 7 (3), 1-11.

Mingers, J., 2001. *Combining IS research methods: towards a pluralist methodology*. Information Systems Research, 12/2001/3, 240-259.

Munro, M. C. and Davis, G. B., 1977. *Determing Management Information Needs: A Comparision of Methods*. MIS Quarterly, 1 (2), 55-67.

Niehaves, B., 2004. *A Framework for Analysing the Epistemological Assumptions of Research Methods*. In Proceedings of the Innovation Through Information Technology. 2004 IRMA International Conference (M. Khosrow-Pour Ed.), 57-60, New Orleans/LA, U.S.A.

Niehaves, B., Klose, K., Knackstedt, R. and Becker, J., 2005. *Epistemological Perspectives on IS-Development – A Consensus-Oriented Approach on Conceptual Modeling.* Accepted for the Second International Workshop on Philosophy and Informatics (WSPI 2005), Ulm, Germany.

Poe, V., 1996. *Building a Data Warehouse for Decision Support*. New Jersey.

Riebel, P., 1979. *Gestaltungsprobleme einer zweckneutralen Grundrechnung*. Zeitschrift für betriebswirtschaftliche Forschung, 31 (1), 863-893.

Rockart, J. F., 1979. *Chief Executives define their own Data Needs*. Harvard Business Review, 30 (Mar-Apr), 81-93.

Sapia, C., Blaschka, M., Höfling, G. and Dinter, B., 1998. *Extending the E/R Model for the Multidimensional Paradigm*. In Proceedings of the International Workshop on Data Warehouse and Data Mining (DWDM'98), 105-116, Singapur, 19.-20. November 1998.

Sethi, V. and Teng, J. T. C., 1988. *Choice of Information Requirements Analysis Method: An Integrated Approach*. INFOR, 26 (1), 1-16.

Szyperski, N., 1980. *Informationsbedarf*. In Handwörterbuch der Organisation (E. Grochla Ed.), 904-913, 2. Ed. Stuttgart.

Tarski, A., 1944. *The Semantic Concept of Truth and the foundation of semantics*. Philosophy and Phenomenological Research, 4/1944, 341-375.

Wedekind, H., 1992. *Datenbanksysteme I. Eine konstruktive Einführung in die Datenverarbeitung in Wirtschaft und Verwaltung.* 2. Ed. Mannheim et al.