

# Metadata-Based Integration of Qualitative and Quantitative Information Resources Approaching Knowledge Management

B. Rieger

A. Kleber

E. von Maur

Institut für Informationsmanagement und Unternehmensführung

Universität Osnabrück

Katharinenstraße 3, D-49069 Osnabrück

www.oec.uos.de/wi2

**Abstract-** This paper presents a concept for the integration of quantitative and qualitative information sources with their accompanying management support functionalities from navigation and retrieval up to analysis and business intelligence. The integration is realized by a common keyword-based metadata base, retrievable and extendible by the end user on a web-based platform. This enables a dynamic acquisition of supplementary information on the usage, usability and benefit of basic and derived information objects, e.g. data warehouses, data marts, OLAP cubes, reports or (textual) documents. Being extended by functions to automatically catch contextual links during system usage, the concept is discussed as a contribution to the implementation of knowledge management. The concept is being developed and successfully tested in the practical environment of a reference project for the implementation of an IT-infrastructure to support decentralized decision-making at a German university.

## MOTIVATION AND RELATED WORK

In the past decade, remarkable progress was reached in management support systems (MSS) research and development [1, 2, 3, 4]. Data warehouses are built to implement an integrated and consistent collection and a supply chain of decision relevant information [5, 6], however, still being restricted to numerical data [7]. The nature of analytical reporting tools, dedicated to data warehouses, like online analytical processing (OLAP), rather consolidates this boundary [8]. At least at the side of end users, supplementary qualitative information needed for effective retrieval, navigation, analysis, and evaluation is still missing [9, 10, 11].

It must be admitted that in most cases a comprehensive metadata subsystem is included, implementing almost all facets demanded [12, 13, 6] like source data types, channels for loading, steps for transformation, and aggregation and so on. Yet, their primary intention (and use) is to manage the (technical) processes of loading and cube generation. If made available to end users at all, the continuous supply chain is broken by copy management or even reentry into another metadata store, proprietary to OLAP-tools for example,

proliferating redundancy and inconsistency. Similar trends can be observed in other areas of research contributing to management support, e.g. document management systems. In most cases, metadata management is done separately neglecting linkage to the other domains, although addressing complementary needs of decision-makers.

A solution might be to extend Devlin's demand for the integration of different types of metadata [6] across boundaries of specific methodologically devided management support systems. The metadata component would have to be outsourced by the MSS-components like data warehouses, business intelligence tools or document management, and repositioned on a superior level. A similar request, going less far in outsourcing, is presented by the concept of an additional layer between information sources and users, called MIS-broker [14]. In any case, as a consequence, a common representation scheme for metadata, being capable to serve the different semantic needs, has to be found. The paradigm of object orientation might best fit these requirements [11, 15]. Similar approaches for conceptually centralized metadata management can be found in the Data Warehouse Quality community [16, 17].

Another important boundary restricting further steps of effective use of management support systems concerns the maintenance of metadata. This links to the question who is authorized for input or change and when update is allowed to be done, i.e. continuously during analytical use for decision-making.

This directly leads to the issues of organizational learning and knowledge management [18]. The classification of knowledge to be extracted in this area [19] shows striking similarities to the extended understanding of metadata derived above, conceptually adding evolutionary dynamics. Despite obvious differences according to intention and content, it seems to be promising to check if the object oriented approach mentioned above will be able to implement one integrated metadata system for both directions. Last but not least, the continuous interactive acquisition of "metadata" inherent to knowledge management has to be transferred to the

end users of all scopes of management support systems. Altogether, this is proposed as a contribution to close the gap in further progress concerning the effectiveness of management support systems. The feasibility in general and the potential benefit is to be justified first by a practical case. This addresses the complaint that new approaches such as knowledge management remain strategic reflection, often omitting concrete implementation phases [18].

#### PROJECT OBJECTIVES

The integrative concept presented is driven by the special characteristics of a research and development project at a German university. The overall objective is to implement an IT-infrastructure to introduce and support decentralized decision-making. The president shall be responsible for strategic decisions (only), e.g. the opening or closing of departments. The chancellor shall be responsible (only) for optimal services on the operational level. The professors are responsible for the effectiveness and efficiency of research and teaching. Furthermore, public authorities and externals shall be included as users, e.g. for reasons of control.

Due to this kind of application area, the ratio of qualitative to quantitative data is extremely high. Lots of qualitative data, primarily available as both structured and unstructured textual documents, e.g. generated by regular evaluation processes, must be made selectively accessible to decision-makers. Communication problems of the past concerning a common understanding (the right meaning) of quantitative measures, e.g. the ratio of third party resources to the number of students, and even the total number of students, have to be prevented by linking qualitative metadata to the quantitative measures and reports. In a decentralized structure of decision-making, these problems in communication are likely to increase. Finally, the increasing communication flows, e.g. to regularly coordinate and evaluate goals and measures, have to be managed in a persistent way.

Without any doubt, this pattern can be transferred to the field of many businesses.

#### LESSONS LEARNED FROM FIRST APPROACH

At first, a classical data warehouse was implemented, filled with lots of quantitative data about human and financial resources, students, programs, graduates etc. Comprehensive metadata about the source databases and the steps of transformation were collected and stored with the tools for extraction, transformation and loading, as usual in the state of the art. Data was made available to decision-makers on different levels of aggregation by reporting and OLAP-tools with web-based clients. Once again, metadata about the steps in calculating and aggregating key indicators were collected and stored together with the OLAP-reports, as usual in the state of the art. Partially, these metadata were made accessible to end users, due to the features of the client tools. Most of them,

however, had to remain invisible to users and are exclusively utilized by the OLAP-tools internally.

In parallel, lots of textual documents, including important qualitative information from evaluation reports, were made accessible by a web-based frontend. These documents were split up into fragments hierarchically and stored into a text database. Metadata about the hierarchical structure of the fragments, their topics and other dimensional aspects, being appropriate for selective retrieval, was analyzed and added to the text database manually. The web-based frontend was extended to partially make use of this kind of keyword-based classification.

Users being confronted with both information sources soon complained about the missing links needed for an integrated retrieval and navigation. The common presentation over the web, as often used, of course cannot accomplish this integration. Furthermore, conflicting interpretations of measures by users were observed, even in the same department. At best, this led to time consuming discussions, and, in many uncovered cases, simply to wrong decisions.

The roots of the problems described can easily be identified in the separated metadata dictionaries, which both can only partially be accessed by decision-makers. However, the integration into one common metadata dictionary alone, is not effective in the long run. The metadata handled this way only covers the knowledge of information providers during the phases of extraction, transformation, loading, and the design of measures as well as reports. Not covered is the knowledge about the decision contexts where the information objects can be applied, neither their effectiveness or benefit nor the (types of) decisions derived. This knowledge is generated dynamically and continuously by the decision-makers during retrieval and navigation in the context of concrete decision situations [20]. This matches exactly the procedural view of knowledge management (organizational learning) [18]. If a learning organization is intended, this virtual knowledge must not be lost but has to be caught as well.

Therefore a concept improving the problems described, which result from separately applying the different lines of information technologies in management support, must address the following three aspects:

1. one common integrated metadata base for all types of information sources,
2. full access to all levels of the metadata base, and
3. extendibility of the metadata base by end users.

#### THE CONCEPT

Fig. 1 illustrates the architecture developed and tested in the project mentioned above. The classical information sources involved are a quantitative database on the left, including a data warehouse and derived views or aggregates for reporting and analysis (data marts, cubes), and a text database for qualitative data (document contents) for text retrieval on the right. Both are currently implemented upon a

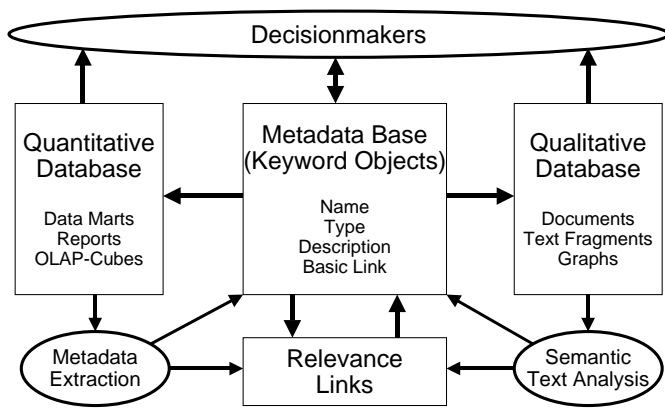


Fig. 1. Conceptual architecture

relational database system (RDBMS) and can be accessed by end users using web-based tools for query, navigation, and analysis (OLAP). Conceptual preparations and a prototype implementation towards object-oriented database technology are done in parallel [11].

In order to realize the requirements derived above a third database is introduced, including keyword objects which can be linked to every component of the quantitative and qualitative databases. This architectural design component is adopted from earlier work addressing the integration of

structured and semi-structured information in another domain [10]. The keyword objects allow multidimensional classification and are best to be implemented in an object-oriented environment. For the time being, it is based on a RDBMS. The contents of this metadata base is also made accessible to end users through the same dynamic webserver as the quantitative and qualitative databases, including functionalities for searching (see below).

The internal structure of the classified keyword objects (see fig. 2) comprises a name, a comprehensive description, a basic type, and, optionally, one basic link to one elementary component of either the quantitative or the qualitative database. In order to model multiple links, in the present relational implementation prototype, a table of relevance is added, including the top-down relevance connections between keyword objects. Thus, any hierarchy or network can be modeled and be made available to end users for drill-down and drill-across navigation.

The keyword objects represent the metadata from each information source included. They are filled both automatically and manually by information providers during the processes of extraction, transformation, loading, and design of measures or (analysis) reports. For the automatic case, procedures have been written to extract metadata [21], e.g. field descriptions, out of the data dictionaries of the source

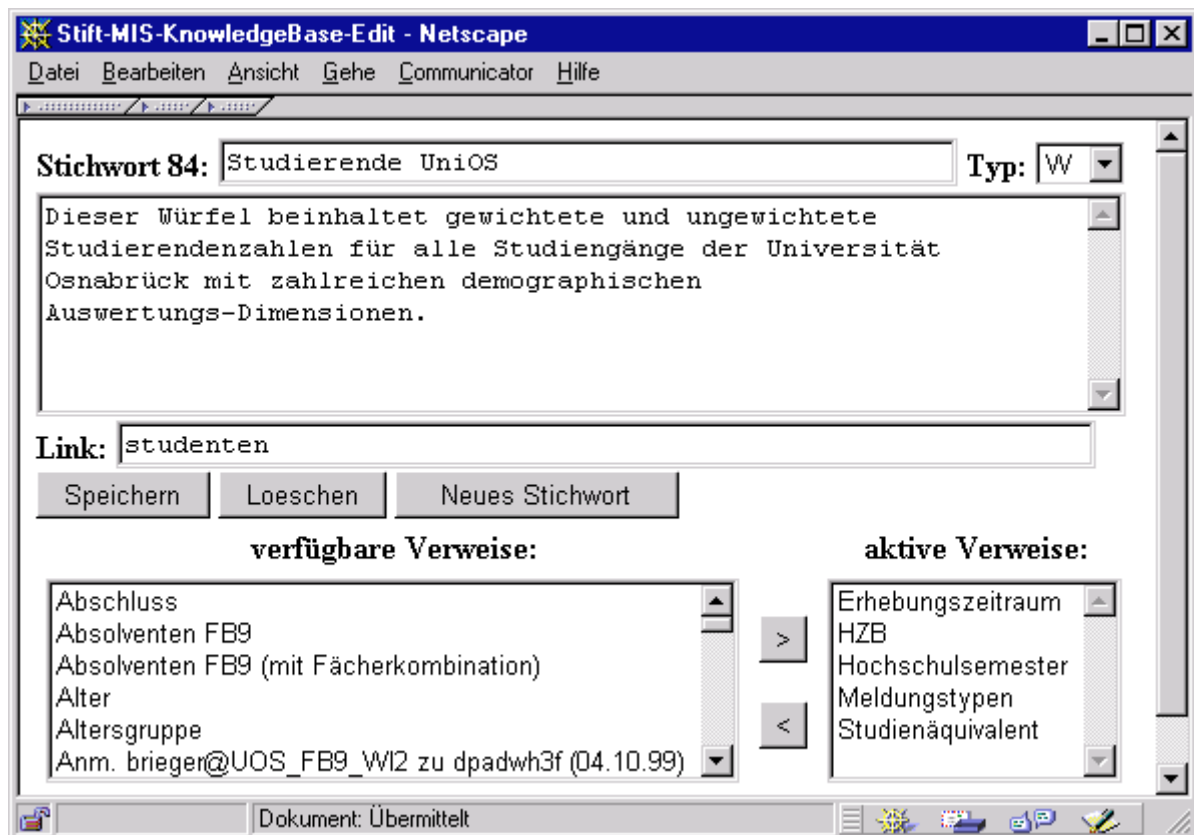


Fig. 2. Editor for keyword objects in the common metadata base

database systems. In case of document bases, knowledge-based techniques for semantic text analysis are to be adopted, which have been successfully implemented for retrieval in scientific domain [22]. In the manual case, information providers can input descriptions, explanations, and usage hints or restrictions for any information objects they generate, both assembling and disassembling. Thus, even reports being composed from many measures and dimensions can be documented. The natural links between the report and the measures or dimensions involved are to be monitored from the design context and added to the relevance table automatically. Of course, additional links can be added manually.

This feature already contributes to the requirement of dynamic knowledge acquisition as essential part of knowledge management. Up to now, however, metadata input is restricted to information providers only. In order to also accompany decision-makers, the metadata warehouse is opened for editing by selected end users on the web-based front end side. Of course, their input is also classified according to the context of use. Once again, this can be done both automatically and manually. One example for an automatic classification is the integration of commenting functionalities of OLAP-clients: a keyword name is generated due to a predefined pattern including the author and the date, for example; the textual comment itself is directly added to the metadata base; a link between the newly added keyword object and the keyword object linking to the report being

commented is generated automatically; further manual linking for classification, e.g. the type of comment or decision situation, can be made available to the end user.

Due to the generic and extendible concept of metadata handling, any type of information about any type of information can be acquired decentrally, administered centrally and delivered instantly. The classical process of navigating through different trees of available information sources in order to search information relevant to concrete decision problems is replaced by a context relevant search request representing the decision problem by the means of keyword objects in the metadata base. As a result, a dynamically generated list of relevant keyword objects with direct links to concrete information objects or indirect links to other keyword objects is presented as a starting point for individual navigation. In the future, the keyword objects for the search request are also to be generated by the means of a semantic knowledge-based analysis of a free text query [23].

#### AN APPLICATION CASE

The following case is taken from the prototype implementation of the MSS-project mentioned above. Fig. 3 shows the matching keyword objects as a result of a query against the metadata base. Each keyword object represents either an information object of the information sources attached or the collection of other keyword objects, thus enabling any contextual information composites.



Fig. 3. Integrated information selection driven by decision context

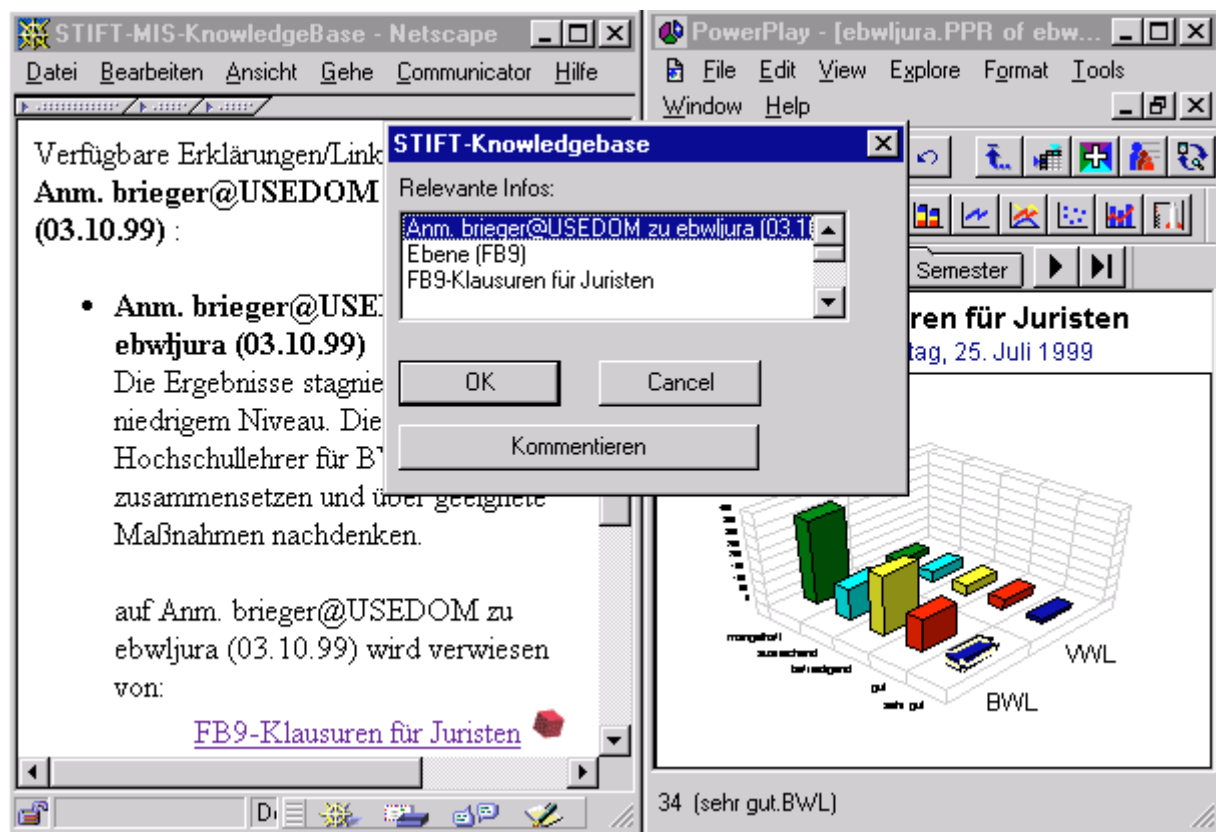


Fig. 4. Cross-boundary navigation driven by relevance

Each keyword object presents itself according to type. The keyword name and textual description are obligatory, author and date are optional. In the case of basic keyword objects, direct links to the information sources, e.g. dynamic queries against a data warehouse, OLAP-cubes, textual or document databases or simply websites are included, with icons indicating their types. Furthermore, the table of relevance links is listed beneath each keyword object in two sections for both drill-down and drill-up navigation. Once again, the type of the relevant keyword object is indicated by icons. Altogether, this comprises all three forms of information sources integration provided by Whinston's MIS-broker [14].

Drilling to basic keyword objects results in activating the corresponding (web) server. The window on the right hand side of fig. 4 shows an example for a keyword object of an OLAP-cube type, triggered by the corresponding keyword link in the drill-up list of relevant keywords in the left window. To enable two-way navigation each client tool for all information sources included has been extended with an interface to the metadata base, realizing a (backwards) drill-through functionality. This is demonstrated in fig. 4 by the pop-up window in the middle, listing all keyword objects available for drill-down in the metadata base for the initiating information object. The selection of a keyword from the list directly leads back to the metadata base system, querying it against the selected keyword. Altogether, this allows a

decision-maker to first navigate in the metadata base system to any basic keyword object, e.g. referencing an OLAP-cube, to switch there, to use the analytic functionality of the OLAP-client, and, finally, to retrieve at any time relevant supplementary information for any component of the cube, available in the metadata base, e.g. the description of a measure or dimension used in the cube or even a textual comment attached to the cube by him or any other user.

So far, this implements static retrieval functionality, only. In order to allow the extension of the metadata base for the acquisition of new insights of decision-makers during (analytical) navigation, the interface mentioned above is supplied with a write function. Thus, end users adopt the role of information providers. A keyword object is generated automatically due to predefined patterns depending on the initiating type of basic information object. For the time being, the pattern comprises author, date, type, and automatic relevance links, so that the contents can be retrieved by other end users immediately. The left window in fig. 4 shows such a keyword object by example.

#### CONCLUDING REMARKS

Beyond the technological solution presented, further organizational and social aspects have to be considered, being critical for the success of the system in the long run. In

addition to literature, these aspects also had to be derived empirically from observing and interviewing key users in the practical case. Neglecting these aspects is about to initiate the collapsing loop of electronic knowledge bases described by Probst [18]: Insufficient quality, lacking evident benefit and reliability of the knowledge base in the beginning and other social barriers restrain a broad and active use. As a consequence, both comfort in access and knowledge base content are going not to be customized as necessary to surpass the natural process of knowledge aging, thus closing the collapsing loop. This also addresses the question if knowledge management is feasible anyway or if knowledge can be managed electronically at all. This discussion is not to be augmented ideologically here. Precautions appropriate for breaking the barriers are to be presented, instead, by additional system components both structurally and procedurally during system introduction and operation. Therefore, the primary focus is to improve management and decision support by the integration of information about the use and usability of quantitative and (other) qualitative data into the process of analysis and decision-making, thus approaching knowledge management.

Mastering insufficient quality is up to the end users. They have to learn that they themselves are responsible for the contents of the system.

The most important reason observed not to use the system as planned, especially to actively input one's own knowledge, lies in the loss of control who else will be able to access the input. A similar problem already exists for components of the quantitative database, not so much for individual but rather for organizational reasons. One possible solution are corresponding access rights usually managed centrally by the information systems department in the past. In an emerging decentralized environment of information providers, who are to be encouraged to spool out information to others, this will not work any longer.

Therefore, the principle of information ownership is going to be introduced in the project mentioned, allowing the information provider solely to determine the scope of accessing people. This principle is applied to all kinds of databases involved, both the referenced databases and the metadata base, as well. During any input, additional specifications of authorized users have to be made explicitly, otherwise restricting access to the provider only, by default. For reasons of comfort and overall consistency, the access rights to be specified are to be taken from one central authorization management system covering the entire organization and all information systems. The user administration of an operating system based mailing server might be a good choice, allowing the specification of individual groups. Once again, another step of outsourcing is needed, concerning the authorization functionalities proprietary to many web-based clients providing information today, e.g. web-based OLAP-clients.

This principle rigorously being implemented, at least encourages the personal use of the system for the administration of one's own individual knowledge, in a first step. Being

put into a context of group decisions, responsibilities and rewards, it is likely to better convince the users of the benefits of further information sharing, step by step. This corresponds to the concept of transferring personal data into public data [12]. Essentially, the end users decide themselves with whom to share which information. Then, possible refusals are not different to the situation without the system and must be solved outside the system on an organizational and social level in any case. The system, however, does not cause additional barriers.

As a by-product, the complexity of access control, growing dramatically with the integration of information sources, is decreasing. Experiences from the project proved very soon that this cannot be managed centrally. In this sense, the question about the feasibility of knowledge management is answered in the negative, if being understood as management of knowledge from the top of the organization. The solution proposed here is to simply install a communication infrastructure, filled with accelerating and balancing feedback loops, persistently gathering and disseminating information about information processing in problemsolving (knowledge) bottom up [24]. Alternatively to knowledge management, this may also be assigned to the second issue of MSS, being defined as the use of related information and communication technologies to support management [25].

## REFERENCES

- [1] St. Alter, *Information systems: a management perspective*, 3. ed., Addison-Wesley, Reading, Mass. et al., 1999.
- [2] C.W. Holsapple and A.B. Whinston, *Decision Support Systems: a knowledge-based approach*, West Publ., Minneapolis/St. Paul et al., 1996.
- [3] E. Turban, *Decision support and expert systems: management support systems*, 4. ed., Prentice Hall, Englewood Cliffs, NJ, 1995.
- [4] H.J. Watson, G. Houdeshel, and R.K. Rainer, *Building executive information systems and other decision support applications*, John Wiley & Sons, New York et al., 1997
- [5] W. Inmon, *Building the Data Warehouse*, 2. ed., John Wiley & Sons, New York et al., 1996.
- [6] B. Devlin, *Data warehouse - from architecture to implementation*, Addison-Wesley, Reading, Mass., 1997.
- [7] R. Kimball, *The data warehouse toolkit*, John Wiley & Sons, New York et al., 1996.
- [8] E. Codd, S. Codd, and C. Salley, "Providing OLAP (On-line Analytical Processing) to user-analysts: an IT mandate", white paper, Arbor Software Corporation, 1993.
- [9] J. Fedorowics, "Document-based decision support", in R.H. Sprague jr., H.J. Watson, *Decision Support for management*, Prentice Hall, New Jersey et al., 1996.
- [10] B. Rieger, K. Brodmann, D. Krüger, E. von Maur and St. Postert, "UniWeb: ein integratives Konzept zur datenbank-gestützten Verwaltung, Navigation und Distribution multimedialer Informationsobjekte", in *Integration externer Informationen in Management Support Systems*, Univ. Dresden, 1998.
- [11] B. Rieger, E. von Maur and St. Postert, "Object Warehouse: Rekonzipierte Entscheidungsbasis des Decision Supports", in *Proc. 2. Workshop "Data Mining und Data Warehousing als Grundlage moderner entscheidungsunterstützender Systeme" (DMDW99)*, LWA99 Sammelband, pp. 97-107, Univ. Magdeburg, 1999.
- [12] B. Devlin and P. Murphy, "An architecture for a business and information system", in *IBM Systems Journal*, vol. 27, no.1, pp. 60-80, 1988.

- [13] W.H. Inmon and R.D. Hackathorn, *Using the data warehouse*, John Wiley & Sons, New York et al., 1994.
- [14] S. Ba, R. Kalakota and A.B. Whinston, "Using client-broker-server architecture for intranet decision support", in *Decision Support Systems*, vol. 19, pp. 171-192, 1997.
- [15] U. Frank, "An object-oriented architecture for knowledge management systems", working paper no. 16, Institut für Wirtschaftsinformatik, Universität Koblenz-Landau, 1999.
- [16] M. Jarke, M.A. Jeusfeld, C. Quix, T. Sellis and P. Vassiliadis, "Metadata and data warehouse quality", in M. Jarke, M. Lenzerini, Y. Vassiliou and P. Vassiliadis, *Fundamentals of data warehouses*. Springer, Berlin, Heidelberg, pp. 123-158, 2000.
- [17] M. Jarke and Y. Vassiliou, "Foundations of data warehouse quality: an overview of the DWQ project", in *Proc. of the 2<sup>nd</sup> International Conference on Information Quality*. Cambridge, MA, pp. 299-313, 1997.
- [18] G. Probst and K. Romhardt, "Bausteine des Wissensmanagements - ein praxisorientierter Ansatz", Cahier de recherche, HEC, Université de Genève, 1998.
- [19] T.H. Davenport, "Some principles of knowledge management", in *Strategy - Management - Competition*, vol. 2, pp. 34-40, 1996.
- [20] S. Ba, K.R. Land, and A.B. Whinston, "Enterprise decision support using Intranet technology", in *Decision Support Systems*, vol. 20, pp. 99-134, 1997.
- [21] B. Rieger and K. Brodmann, "Mastering time variances of dimension tables in the data warehouse", unpublished.
- [22] K.-U. Carstensen, B. Diekmann, G. Möller, "GERHARD (German Harvest Automated Retrieval and Directory)", unpublished.
- [23] M. Ronthaler, "Osiris: Qualitative Fortschritte bei der Literaturrecherche", in J. Dassow and R. Kruse (Hrsg.), *Informatik '98: Informatik zwischen Bild und Sprache*, Springer, Berlin et al., 1998.
- [24] P.M. Senge, *The fifth discipline*, Doubleday Books, New York et al., 1990.
- [25] M. Scott Morton, "State of the art of research in Management Support Systems", CISR-MIT, working paper no. 107, 1983.