A chronological study of paradigms for data warehouse design

Estudio cronológico de paradigmas para el diseño de almacenes de datos

A. Cravero[1], S. Sepúlveda[2]

ABSTRACT

Data warehouses must homogenise and integrate data from an organisation's areas to extract relevant knowledge for orientating decision-making. This is not an easy task, which is why several approaches have been developed which can be classified according to how information is obtained: supply-driven, demand and a hybrid of the first two. This study offers a conceptual framework and a chronological study of the approaches adopted according to the paradigm used, providing a comprehensive context for better understanding the current state of data warehouses and their evolution.

Keywords: Data warehouse, paradigm, chronological study.

RESUMEN

Los almacenes de datos deben homogeneizar e integrar la información de las diversas áreas de una organización con el fin de extraer conocimiento relevante para la toma de decisiones. Su desarrollo no es tarea fácil, por lo cual se han desarrollado varios enfoques que se pueden clasificar de acuerdo con la forma como se obtienen los requisitos de información: impulsados por la oferta, por la demanda, e híbridos. Este estudio ofrece un marco conceptual y un estudio cronológico de los estos enfoques conforme al paradigma utilizado, lo que proporciona un amplio marco para comprender de mejor manera el estado actual en el área de los almacenes de datos, así como su evolución.

Palabras clave: almacén de datos, paradigma, estudio cronológico.

Received: January 5th 2011
Accepted: April 30th 2012

Introduction

Data warehouses (DW) are designed to support organisations' decision-making (Giorgini *et al*., 2008; Mazón, 2005; Rizzi *et al*., 2006); these systems must homogenise and integrate data from an organisation's different areas into a large data repository so as to take advantage of such unique and detailed representation, thereby enabling the extraction of knowledge which is relevant to decision-making. In themselves DW represent a large data repository that does not say much; as with operational databases, additional tools are needed to consult and analyse the stored data. Lacking the right tool, one cannot extract knowledge from a DW which is valuable to an organisation and the entire system will fail in its aim of providing information which supports decision-making. Online analytical processing (OLAP) (Codd *et al*., 1993) is a tool which facilitates analysing information and navigation throughout an entire DW to extract relevant knowledge for such organisation.

Developing a DW is not an easy task and some interesting difficulties arise when attempting to do so. Nowadays, although a standard model is still lacking, it is widely assumed that DW design must follow a multidimensional paradigm (Kimball and Ross, 2002). As a result, a lot of effort has been spent in creating methodologies and approaches enabling the creation of a multidimensional model of a DW (Romero and Abello, 2009).

According to Winter and Strauch (Winter and Strauch, 2003), multidimensional model approaches can be classified according to DW requirements: demand-driven, supply and hybrid approaches seeking to combine the first two. This study offers a conceptual framework and a chronological study of approaches for designing DW according to the paradigm used, providing a comprehensive context for better understanding the current state of DW and their evolution.

A literature review was made to obtain a pertinent set of DW design approaches; a set of key words was defined which were related to DW design, modelling and representation. Forty-three studies were collected, of which 15 were selected, dating from 1998 to 2010. Studies which did not provide a model for designing DW were not considered and when more than one study

[1] Ania Cravero Leal. *Affiliation:* Universidad de la Frontera, Chile. ScD. in computern information systems, Atlantic International University, United States. MSc. In Information technologies applied in bussines, Universidad Politécnica de MAdrid, España. E-mail: acravero@ufro.cl

[2] Samuel Sepúlveda Cuevas. Affiliation: Universidad de la Frontera. ScD. Canidate in Informatics Application, Universidad de Alicante, España. MSc. in direction and management of systems and TIC. Universidad oberta da Catalunya, España. E-mail: ssepulve@ufro.cl

How to cite: A. Cravero, Sepúlveda S. (2012). A chronological study of paradigms for data warehouse design. Ingeniería e Investigación. Vol. 32, No. 2, August 2012, pp. 58-62.

had been published on the subject, the first authors to present such methodology was selected.

Basic concepts

The classical definition of DW was coined by Inmon (Inmon, 1996) as being a collection of historical, subject-orientated, non-volatile, integrated data which has been designed to support an organisation's decision-making.

More than just being a simple data summary, a DW may be defined in three stages: extracting data from different data sources, the consistent transformation and uploading of data into a DW (Abril and Pérez, 2007) and efficient and flexible access to integrated data. The first two stages are known as extract-transform-load (ETL).

A DW's main contribution lies in its ability to convert data into strategic information, thereby supporting decision-making at an organisation's highest levels. This ability is supported by the OLAP tool (Codd *et al*., 1993) which provides end-users with configurable views of data from different angles and at different aggregation levels (Silva-Paim and F.B.-Castro, 2003).

Data is organised multi-dimensionally (known as star schema) for rapid, flexible OLAP consultations, information being classified according to facts and dimensions (Kimball and Ross, 2002). Facts are numeric data or data representing a specific industrial activity to be analysed; dimensions are individual perspectives of data determining the granularity (detailed data) adopted for representing a fact. Units of facts and their values are called measurements (Kimball and Ross, 2002); Figure 1 illustrates the complete process.

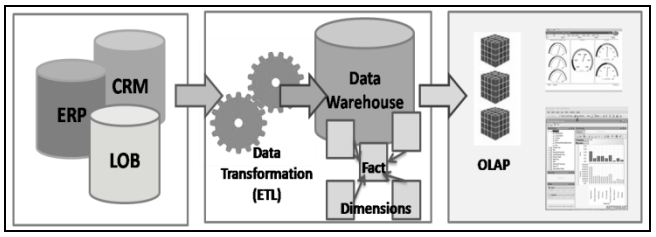

Figure 1. Extracting information from a DW

Paradigms for data warehouse design

This section describes basic DW design paradigm characteristics. A reader can find a different classification in (List *et al*., 2002), where a link is established between design methodology and requirement domain.

Supply–driven

Supply-driven approaches (also known as data) initiate DW modelling from a detailed analysis of data sources to determine the elements (facts and dimensions) which may be useful in producing analysis supporting decision-making in a particular business. The information stored in the facts usually represents business measurements (e.g. how many products are being sold? How many patients are being treated? How long does a given process take? etc.) and the dimensions represent the framework for analysing such measurements (e.g. time, customer or product) (Winter and Strauch, 2003).

Demand–driven

Demand-driven approaches (also known as requirement-driven or goal-orientated) concentrate on determining users' needs (as is normally done in information systems) to then create a multi-dimensional DW design according to selected goals.

Some authors (Anwer and Ikram, 2006; Bubenko, 1993; Jacobson, 1995; Mylopoulos, 2006; Yu and Mylopoulos, 1994) have stated that today's systems are viewed in terms of providing a company with solutions to specific problems; the relationship between systems and their surroundings are therefore expressed in terms of goal-based relationships. This is partly motivated by the greater dynamism present in businesses and the organisational environment, where systems are increasingly used for fundamentally changing business processes instead of automating information flow practice. Modelling techniques must support the "why" rather than the "how" to produce new types of analysis (e.g. to analyse whether it is possible to fulfil an objective with the activities being planned) (Yu and Mylopoulos, 1998).

Hybrid

A hybrid approach seeks to combine both the above to design a DW from data sources, but taking end-users' needs into account. The main difference is that this type of approach can intersperse supply- and demand-driven approaches in carrying out each stage of DW development, benefitting from information collected throughout the whole process (Romero and Abello, 2009).

Multidimensional data warehouse design approaches

This section describes some approaches known for their multi-dimensional DW design; each approach has been classified according to the paradigm used.

Supply–driven

Golfarelli and Rizzi (Golfarelli *et al*., 1998; Golfarelli and Rizzi, 1998) made a formal, structured overall study of multidimensional design; it was partially automatic, consisting of six well-defined stages. Their first step is emphasised since this is where an information system is analysed and a conceptual scheme produced (i.e. an entity-relationship (ER) diagram) or a logical (i.e. relational) Schema.

Moody and Kortink (Moody, 2003; Moody and Kortink, 2000) proposed a methodology for developing multidimensional schemes via an ER model. This was one of the first supply-driven approaches introduced into the literature. Although they were not the first to use an ER model, they presented a structured, formal methodology for developing logical schema.

Phipps and Davis (Phipps and Davis, 2002) proposed one of the first approaches for automating part of the design process. Their main aim was automating a supply-driven process based on two basic premises: numerical fields represent measurements and numerical fields plus a relational table probably assume a factual role within AD. Also any table related to one-to-many cardinality would likely be able to play a dimensional role. The authors provided a detailed pseudo-algorithm and automation would thereby have been immediate. However, this approach produced too many results, and demand-based scenarios had to be created to filter the results according to end-user requirements.

Vrdoljak *et al*., (Vrdoljak *et al*., 2003; Vrdoljak *et al*., 2006) have presented a semi-automatic supply-based approach for obtaining

logical models by considering XML schemata as a source considered a de facto standard for exchanging semi-structured data.

Demand–driven

Kimball *et al*., (Kimball *et al*., 1998) presented the multidimensional model as we know it today. Being the first approach to obtain a multidimensional schema, it did not present an explicit design procedure, but only a detailed guide for identifying multidimensional elements to generate multidimensional schema. The methodology's presentation was quite informal and was based on examples rather than explicit rules. Data sources were not considered and it was scarcely suggested that data sources be looked at to find potential data marts of interest.

Winter and Strauch (Winter and Strauch, 2003) presented a design methodology developed by analysing several DW projects created in participating companies. Their approach was quite different from the rest of the methodologies reviewed, arguing that no specific data model need be chosen to express the conceptual schema developed. Selecting a data model to use (one in the multidimensional model), they created a step-by-step guide regardless of such data, identifying the best practices that a DW design project must include according to prior analysis.

Giorgini *et al*., (Giorgini *et al*., 2005) offered a demand-driven approach; however, their approach was a hybrid if one considers reconciliation data sources with the requirements obtained from the business strategy. The authors introduced a methodology orientated towards an agent based within i* framework (Yu, 1995). They argued that it was important to consider the model of an organisation and models of each actor's decision-making to capture a DW's functional and non-functional requirements . Later, they introduced the Tropos objectives model (Bresciani *et al*., 2004) for the requirements-specifying stage (GRAnD approach) (Giorgini *et al*., 2008).

Prakash and Gosain (Prakash and Gosain, 2008) focused on the widest context possible for an organisation's goals when designing a DW. Strategic goals then enable identifying the relevant set of decisions, in turn aiding determining DW content. They used the information scenario to gain a technical point of view; this is a typical interaction between a DW and the decision-maker and consists of a sequence of pairs: i.e. information request / response. An information request is formulated as a statement in an adapted form of SQL called SQL specification; the responses represent the information possibilities which can obtained from a set of decisions selected for a determined goal.

Hybrid

Cabibbo and Torlone (Cabibbo and Torlone, 1998; Cabibbo and Torlone, 1999) presented an approach involving logical schema by using ER diagrams. It could also produce a multidimensional schema in terms of relational databases or multidimensional matrices. The multidimensional elements had to be manually identified by a user, therefore from the requirements. Just as with Kimball's approach, this approach was quite informal. Nevertheless, these methodologies formed the basis for all the other methodologies.

Böehnlein and Ulbrich-vom (Böhnlein and Ulbrich-vom, 2000) presented a hybrid approach for deriving a logical schema in a structured entity relationship (SER) diagram. SER is an extension of ER visualising whether there might be dependencies between objects. This is why the authors argued that SER was a better alternative for identifying multidimensional structures.

Bonifati *et al*., (Bonifati *et al*., 2001) presented a semi-automatic hybrid approach consisting of three basic steps: a demand-based stage (using the goal-question-metric (GQM) model), a supply scenario-based stage and an integration stage. The last step aimed at integrating and reconciling both paradigms and generating a viable solution that best reflected a user's needs. This method gave a logical multidimensional schema and was the first to introduce a formal hybrid approach having an integration step reconciling both paradigms. This methodology has been applied and validated in a real case study.

Romero and Abelló (Romero and Abelló, 2006) presented a method for obtaining multidimensional conceptual models from the requirements expressed in SQL queries and relational models. This approach is totally automatic; it was the first to automate end-user needs. Unlike other hybrids, this approach involved supply- and demand-driven approaches simultaneously, benefitting from the comments returned by each paradigm; the approach was thus able to obtain much more valuable data than when both stages were implemented sequentially.

Mazón *et al*., (Mazón *et al*., 2007) presented a semi-automatic hybrid approach in which they first obtained a conceptual DW model from users' needs (goals which must be achieved), then verified whether it were possible to collect data to analyse goals selected from data sources. This involved using query view transformations (QVT) allowing a multidimensional DW model to be automatically generated from users' requirements expressed within the i* framework (Yu, 1995). Later they introduced the business motivation model (BMM) (BRG, 2007) in the requirements-gathering stage for achieving goals which must be considered in the i* model. DW could thus be aligned with a specific business strategy (Cravero *et al*., 2010).

Chronological study of approaches

Figure 2 provides a chronological illustration of each approach described in the previous section.

The last approach for developing supply-driven DW was created in 2004; since then, this type has no longer been considered by researchers, mainly because it has been shown that 80% of DW fail since an organisation's true strategic needs are not represented (Frolick and Winter, 2003; Weir *et al*., 2003; Wixom and Watson, 2001). Weir concluded that it was necessary to align DW objectives with business strategy goals to reduce the failure rate; however, these needs have scarcely been considered by new approaches, leading researchers to concentrate on fulfilling budgets and deadlines, and not on organisational strategy (Weir *et al*., 2003).

As far as demand-driven approaches are concerned, the tendency has been to use goals modelling decisions and scenarios.

By contrast, the ER model was used during the first few years to represent DW in the hybrid approach. However, Bonifati changed the modelling type since 2002 to include user objectives by using the GQM model (Bonifati et al., 2001). This model marks a clear tendency in this type of paradigm since other authors currently use other models like i* and Tropos. It should be emphasised that no standard has emerged in using goal models, as each approach uses something different from the others, including i*, Kaos and GQM (Kavakli, 2002). Automation is much more obvious in this paradigm, even going so far as to develop totally automatic processes. One characteristic worthy of note is the use of transformations between models using standards like

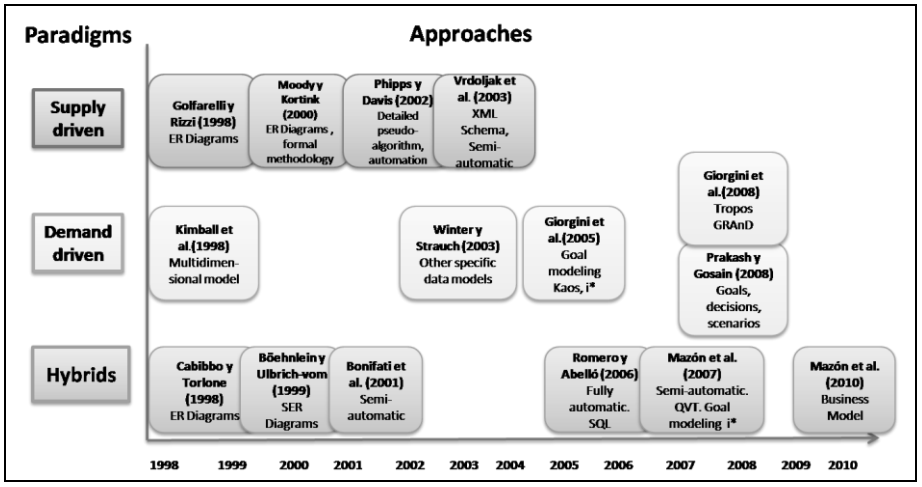

Figure 2. Chronological schema of approaches for multidimensional DW design

QVT; this has been due to a clear representation of different models using an abstraction on three levels: conceptual, logical and physical. As of 2010, the business strategy model has come into use as a source of objectives which must be considered for designing DW.

Despite the approach proposed by Prakash and Giorgini being demand-driven, the trend seems to be towards adopting hybrids as new approaches for developing DW since they lead to confirming that data sources have the necessary information for responding to the strategic questions which company decision-makers must analyse.

Conclusions

This work has presented a conceptual framework and chronological study of demand-driven, supply-driven and hybrid approaches for designing paradigm-based DW.

Some criteria have been introduced for establishing a basis for discussion and detecting tendencies, such as common characteristics or trends applying to such approaches: modelling type or technique used to represent DW, type of model used to represent users' needs and automation.

This study has provided an ample framework for better understanding the current state of DW design and its evolution. The trend in DW design would seem to be towards using hybrid approaches considering the strategy model, using objective models for DW and transformations between conceptual, logical and physical models for achieving automation.

Future work will involve a comparative study of the main approaches and proposing a framework for assessing their maturity level.

Acknowledgement

This work was financed by the Universidad de La Frontera's Research and Post-Graduate Studies' Research Office through DIUFRO Research Project #DI11-0004.

References

Abril, D.,Pérez, J., Estado actual de las tecnologías de bodega de datos y OLAP aplicadas a bases de datos espaciales. Ingeniería e Investigación., Vol.27, N°.1, 2007, pp. 1-12.

Anwer, S.,Ikram, N., Goal Oriented Requirement Engineering: A Critical Study of Techniques. APSEC '06 Proceedings of the XIII Asia Pacific Software Engineering Conference., 2006.

Böhnlein, M.,Ulbrich-vom, A., Business Process Oriented Development of Data Warehouse Structures. Proceedings of Data Warehousing., Physica Verlag., 2000.

Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A.,Paraboschi, F., Designing Data Marts for Data Warehouses. ACM Transactions on Software Engineering and Methodology., Vol.10, N°.4, 2001, pp. 452–483.

Bresciani, P., Giorgini, P., Giunchiglia, F., Iopoulos, J.M.-.Perini, A., Tropos: An agent-oriented software development methodology. Journal of Autonomous Agents and Multi-Agent Systems, Vol.8, N°.3, 2004, pp. 203-236.

BRG, The Business Motivation Model. Business Governance in a Volatile World. www.BusinessRulesGroup.org, 2007.

Bubenko, J.A., Extending the Scope of Information Modeling. Proc. 4th Int. Workshop on the Deductive Approach to Information Systems and Databases, Lloret-Costa Brava, Catalonia., 1993, pp. 73-98.

Cabibbo, L.,Torlone, R., A Logical Approach to Multidimensional Databases. In H. Schek, F. Saltor, I. Ramos, G. Alonso (Eds.). Proceedings of 6th International Conference on Extending Database Technology., Vol.1377, 1998, pp. 183-197.

Cabibbo, L.,Torlone, R., Querying multidimensional databases. Journal Database Programming Languages., 1999, pp. 319-335.

Codd, E., Codd, S.,Salley, C., Providing OLAP to user-analysts: An IT mandate., E. F. Codd and Associates. Vol.32, 1993.

Cravero, A., Mazón, J.-N.,Trujillo, J., Guia de diseño del Almacén de Datos para mejorar el alineamiento de objetivos mediante BMM. Jornadas de Ingeniría de Software y Bases de Datos. JISBD'10. Valencia, España, 2010.

Frolick, M.,Winter, K., Critical Factors for Data Warehouse Failure. Business Intelligence Journal http://www.tdwi.org/research/display.aspx?ID=6592., Vol.8, N°.1, 2003, pp. 61-71.

Giorgini, P., Rizzi, S.,Garzetti, M., Goal oriented requirement analysis for data warehouse design. DOLAP'05, Bremen, Germany, nov., Vol.45, 2005, pp. 47-56.

Giorgini, P., Rizzi, S.,Garzetti, M., GRAnD: A goal-oriented approach to requirement analysis in data warehouses. Decision Support Systems., Vol.45, 2008, pp. 4 - 21.

Golfarelli, M., Maio, D.,Rizzi, S., Conceptual Design of Data Warehouses from E/R Schemes. System Sciences., 1998.

Golfarelli, M.,Rizzi, S., A methodological framework for data warehouse design. Proceedings DOLAP'98., 1998, pp. 3-9.

Inmon, W., Building the Data Warehouse. (2nd ed.). New York etc.: John Wiley & Sons, 1996.

Jacobson, I., Modeling with use cases-Formalizing use-case modelling. Journal of Object-Oriented Programming., 1995.

Kavakli, E., Goal Oriented Requirements Engineering: A Unifying Framework. Requirements Engineering., Vol.6, N°.4, 2002, pp. 237-251.

Kimball, R., Reeves, L., Thornthwaite, W.,Ross, M., The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses. John Wiley & Sons, Inc., 1998.

Kimball, R.,Ross, M., The Data Warehouse Toolkit, second edition, John Wiley & Sons. 2002.

List, B., Bruckner, R.M., Machaczek, K.,Schiefer, J., A Comparison of Data Warehouse Development Methodologies Case Study of the Process Warehouse. In A. Hameurlain, R. Cicchetti, R. Traunmüller (Eds.) Proceedings of 13th International Conference on Database and Expert Systems Applications. Lecture Notes in Computer Science. Aix-en-Provence., Vol.2453, 2002, pp. 203-215.

Mazón, J.N., Pardillo, J.,Trujillo, J., A Model Driven Goal Oriented Requirement Engineering Approach for Data Warehouses. Advances in Conceptual Modeling--Foundations and Applications, Springer, 2007, pp. 255-264.

Mazón, J.N., Trujillo, J., Serrano, M., Piattini, M., Designing Data Warehouses: From Business Requirement Analysis to Multidimensional Modeling. REBNITA'05, 2005, pp. 44-53.

Moody, D.-L., From ER Models to Dimensional Models: Bridging the Gap between OLTP and OLAP Design. Journal of Business Intelligence., The Data Warehouse Institute, Vol.8, 2003, pp. 7-24.

Moody, D.L.,Kortink, M., From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design. In M. A. Jeusfeld, H. Shu, M. Staudt, G. Vossen (Eds.). Proceedings of 2nd International Workshop on Design and Management of Data Warehouses; pp. Stockholm, Sweden: CEURWS. org., 2000, pp. 6.

Mylopoulos, Goal-Oriented Requirements Engineering Part II. Proceedings of the 14th IEEE International Requirements Engineering Conference., IEEE Computer Society, 2006, pp. 4.

Phipps, C.,Davis, K.C., Automating Data Warehouse Conceptual Schema Design and Evaluation. Proceedings of 4th International Workshop on Design and Management of Data Warehouses., Vol.58, 2002, pp. 23-32.

Prakash, N.,Gosain, A., An approach to engineering the requirements of data warehouses. Requirements Engineering., Springer, Vol.13, N°.1, 2008, pp. 49-72.

Rizzi, S., Abelló, A., Lechtenbörger, J.,Trujillo, J., Research in Data Warehouse Modeling and Design: Dead or Alive? Proceedings of the 9th ACM international workshop on Data warehousing and OLAP. Arlington, Virginia, USA, ACM, 2006, pp. 3-10.

Romero, O.,Abello, A., A survey of Multidimensional Modeling Methodologies. International Journal of Data Warehousing & Mining., Citeseer, Vol.5, N°.2, 2009, pp. 1-23.

Romero, O.,Abelló, A., Multidimensional Design by Examples. Proceedings of 8th International Conference on Data Warehousing and Knowledge Discovery., Springer, 2006, pp. 85-94.

Silva-Paim, F.-R.,F.B.-Castro, J., DWARF: AN Approach for require ments Definition and Management of Data Warehouse Systems. RE'03, 2003.

Vrdoljak, B., Banek, M.,Rizzi, S., Designing Web Warehouses from XML Schemas. Proceedings of 5th International Conference on data Warehousing and Knowledge Discovery., Springer, Vol.2737, 2003, pp. 89-98

Vrdoljak, B., Banek, M.,Skocir, Z., Integrating XML sources into a data warehouse. Data Engineering Issues in E-Commerce., 2006, pp. 133-142.

Weir, R., Peng, T.,Kerridge, J., Best practice for implementing a data warehouse: A review for strategic alignment. School of Computing, Napier University, 10 Colinton Road, Edinburgh EH10 5DT UK., 2003.

Winter, R.,Strauch, B., A method for demand driven information requirements analysis in data warehousing projects., Proceedings of the 36th Annual Hawaii International Conference on. System Sciences, 2003, pp. 9-14.

Wixom, B.,Watson, H., An Empirical Investigation of the Factors Affecting Data Warehousing Success. MIS Quarterly., JSTOR, Vol.25, N°.1, 2001, pp. 17-41.

Yu, E., Modelling Strategic Relationships for Process Reengineering., PhD thesis, Computer Science Department, University of Toronto, Toronto (Canada). PhD thesis, also appears as Technical Report DKBS-TR-94-6, December 1994, 1995.

Yu, E.,Mylopoulos, J., From E-R to A-R -- Modelling Strategic Actor Relationships for Business Process Reengineering. Entity-Relationship Approach (ER'94) -- Business Modelling and Re-Engineering, Lecture Notes in Computer Science no. 881 Springer-Verlag, 1994, pp. 548-565.

Yu, E.,Mylopoulos, J., Why Goal Oriented Requirements Engineering? Proc. 4th Int. Workshop on Requirements Engineering for Software Quality, Foundations of Software, Quality – REFSQ., Pisa, Presses Universitaires de Namur., 1998, pp. 15-22.