The Use of Soft Systems Methodology for the Development of Data Warehouses

Roelien Goede
School of Information Technology, North-West University
Vanderbijlpark, 1900, South Africa

ABSTRACT

When making strategic business decisions, managers in large corporations rely on data that is integrated from different sources in their corporation into data warehouse systems. The development of data warehouses entails the integration of data from sources with often conflicting technical infrastructure combined with a clear understanding of the vision of the corporation. The work of practitioners Inmon and Kimball are frequently used as development methodologies. Yet many data warehouse developers do not understand the business objectives. The soft systems methodology provides a set of methods, originating from experience in management situations, for purposeful activity in an organisation. This paper demonstrates how the current data warehousing methodology of Kimball can be enriched by incorporating ideas from Checkland's soft systems methodology.

Keywords: Data warehousing, dimensional modeling, soft systems methodology and systems thinking.

1. INTRODUCTION

Data warehouses are used to provide information for managers when making strategic decisions in an organisation [1]. The development of a data warehouse is often viewed from two perspectives: the back room where the technical data integration is done and the front room where the end-user applications are developed [2], as indicated on Figure 1. The data warehouse team therefor has to interact on two levels with other members of the organisation. First, on a technical level with data owners and secondly, on a business level with end-users who are typically managers. The technical information technology (IT) and the business users have very different viewpoints on the operation of the data warehouse. The systems approach evolved from the need to have a more holistic understanding of a problem situation as a reaction to the reductionist approach often followed to divide and conquer problems [3]. Checkland [4] developed the soft systems methodology (SMM) as guidelines to better understand different viewpoints in a problem situation in order to take purposeful action that is culturally feasible. This paper aims to demonstrate how the soft systems methodology can complement traditional data warehousing development methodologies.

2. METHODOLOGY

This paper reports on a conceptual comparison of data warehousing development methodologies and soft systems thinking and more specifically the SSM. The main section of the paper begins with a discussion of data warehouse development methodology in section 3. The purpose is to provide enough background information on data warehousing to demonstrate the holistic effort required to be successful in providing business users with useful information.

.

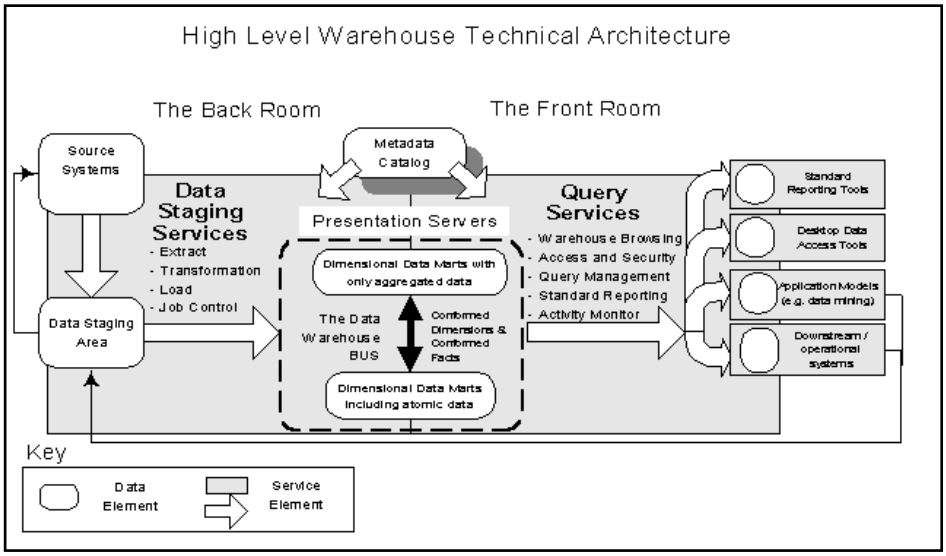

Figure 1 Data warehouse development in terms of back room and front room activities [2]

Section 4 provides an overview of soft systems thinking, focussing on the soft systems methodology (SSM). The purpose of this section is to provide enough information to demonstrate the main ideas of SSM on model building and participative change. Traditional data warehousing methodology is viewed in section 5 from a SSM approach. A conceptual link is made between data warehouse development methodologies and the SSM. This is done in order to develop guidelines in section 6 for the use of the SSM in data warehouse development. The paper concludes with a summary and recommendations for future work in section 7.

3. DATA WAREHOUSE DEVELOPMENT

Larger corporations have a need to integrate data from different sources for use in their strategic decision making. These source systems, often called legacy systems, often function on different infrastructure platforms with different technical data formats [2]. A data warehouse is used to integrate data from different source systems to provide information to business users. Inmon, generally accepted as the "father" of data warehousing, describes a data warehouse "as a subject oriented integrated, non-volatile, and time variant collection of data in support of management decisions." [5]. Reference [6] explains each of the parts of this definition: "Subject oriented: A data warehouse is organised around the key subjects (or high level entities) of the enterprise. Major subjects may include customers, patients, students, and products. Integrated: The data housed in the data warehouse is defined using consistent naming conventions, formats, encoding structures, and related characteristics. Time-variant: Data in the data warehouse contains a time dimension so that it may be used as a historical record of the business. Non-volatile: Data in the data warehouse is loaded and refreshed from operational systems, but cannot be updated by end-users." Kimball simply defines a data warehouse as "the queryable source of data in the enterprise." [7]. The Inmon definition and the explanation of terms aid understanding of the differences between data warehouses and general everyday online transactional information systems.

There are two main methodologies in data warehouse development [1]. Inmon advocates a data driven approach whereby all available data are gathered in the organisation, integrated, and presented in a format that is usable to business users [5]. Inmon argues that the requirements will evolve from the availability of the data in subject-oriented data marts. Kimball on the other hand proposes a requirements-driven methodology based on the collection of user requirement from business users [2]. This paper provides guidelines for the use of SSM in the methodology of Kimball, since it is a user centred approach.

Kimball depicts the lifecycle of a data warehouse project in terms of three main streams of development as indicated on Figure 2. Kimball proposes that the readiness of the organization to start a data warehousing project must be investigated. Two key features are investigated in this regard. Firstly there must be a compelling business motivation for the data warehouse [7]. In other words, there should be a business problem to provide information for. Secondly there must be a business sponsor from the management team of the organisation. Data warehouses that are initiated and motivated from the IT department are seldom successful [7].

Figure 2 indicates three tracks in the development of a DW. The technical track comprises the planning and selection of architecture and infrastructure for the technical operation of the data warehouse. The middle track is most interesting for purposes of this paper as it involves the modeling of the data warehouse from user requirements. The bottom track representing end-user application development is not that different from general application development. Checkland and Holwell [8] provide an in-depth discussion of the use of SSM in general application development.

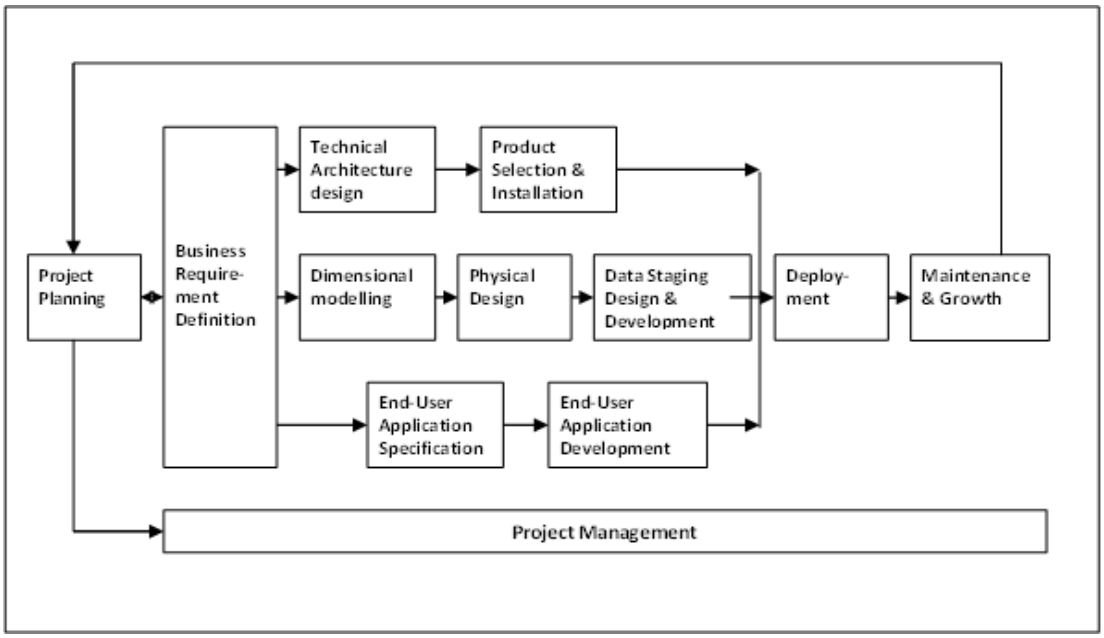

Figure 2 Kimball's business lifecycle of a data warehouse [2]

	Date	Raw Material	Supplier	Plant	Product	Shipper	Warehouse	Customer	Sales Rep	Promotion Deal
Raw Material Purchasing	X	X	X	X		X				
Raw Material Delivery	X	X	X	X		X				
Raw Material Inventory	X	X	X	X						
Bill of Materials	X	X		X	X					
Manufacturing	X	X	X	X	X					
Shipping to Warehouse	X			X	X	X	X			
Finished Goods Inventory	X			X			X			
Customer Orders	X				X	X		X	X	X
Shipping to Customer	X				X	X	X	X	X	X
Invoicing	X				X		X	X	X	X
Payments	X				X			X	X	X
Returns	X				X	X		X	X	X

Figure 3 Data requirements by different business processes in the organization [2]

After collection of user requirement with traditional interview methods the data modeling team can build dimensional data models that provide end users with the information required. Kimball provides a four step model for creating these models [2]. The first step is to select the business process to be modeled. Kimball promotes a holistic understanding of the organisation in terms of business processes and data entity requirements. He uses the bus architecture matrix depicted in Figure 3 to gain an understanding of the data needs of the organisation [2]. The second step is to decide on the level of transaction detail to be modeled in the dimensional model, referred to as the grain of the data. The third step is to identify the dimension tables. Dimension tables hold the descriptive data regarding data entities. Identification involves answering of the "W" questions: What, Where, When, Why and Who. The final step is to add transactional data to the fact table linking all the dimensional data to the specific transaction or event modeled. Numeric values associated with the event, such as the item price of a product are stored in the fact table. Figure 4 provides an example of a simple dimensional model of a retail transaction.

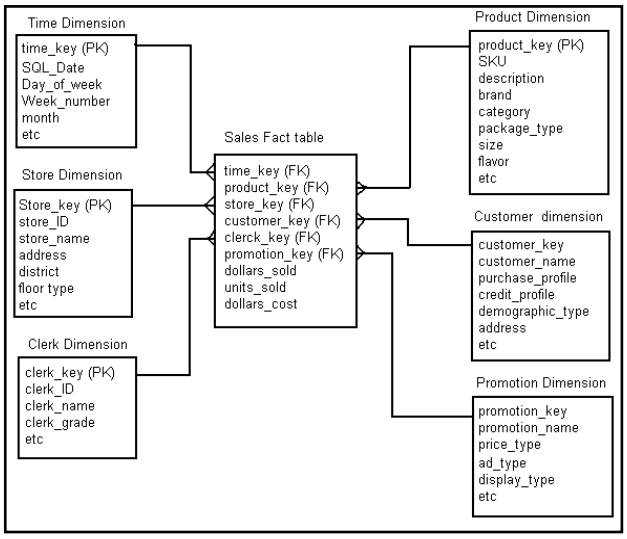

Figure 4 A star schema for a retail sales process.[2]

Different business processes are modeled individually as star schemas such as Figure 4. Dimensions are shared across different star schemas and needs to be standardized (conformed [2]). Figure 5 from [2] depicts this holistic view of the dimensional models (data marts) in an organisation.

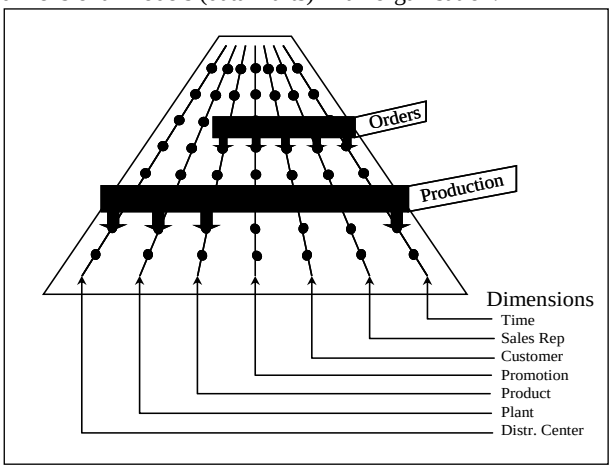

Figure 5 A holistic view of the data used in an organisation [2]

After completion of the modeling process, the data base is designed and data is loaded from source systems of the organisation. The process is called Extract, Transform, and Load (ETL). As indicated on Figure 2, the data warehouse is then deployed and phases of growth and maintenance are started.

4. SYSTEMS THINKING AND THE SOFT SYSTEMS METHODOLOGY

Systems thinking developed as a reaction to the reductionist approach of management science in the period around World War II when management problems were identified and solved using mathematical models. A system is a set of interrelated components or subsystems that work together to achieve a goal [3]. It has emergent properties which are not identifiable in the subsystems and it has built-in control mechanisms to ensure effective achievement of the goal [9]. The environment of the system is the constraints in which it has to function [3]. Management problems are part of problematic situations,

influenced by many complex social factors and therefor difficult to approach by mathematical models alone. Checkland argues that a soft systems thinker views a problem situation as a "mess" and uses systems to make sense of the situation [9]. This is in contrast to early hard systems thinkers who view the problem situation as a group of systems working together. For the soft systems thinker the system is a method of understanding the situation from a specific worldview.

Checkland developed the soft systems methodology (SSM) as set of guidelines to understand a problem situation from different world views and to guide purposeful action to improve the situation [4]. A concise explanation can be found in reference [10]. The following discussion provides a summary of the methodology but does not do the depth of SSM justice. A simplified flow of the SSM is depicted in Figure 6 [10]. A real world situation exist where there are problems which somebody wants to address. Models are developed representing purposeful activity systems of different worldviews in the situation. Each module (depicted by a square shape in figure 6) represents a different worldview. The modeling process will be discussed in the next paragraph. The modules are not descriptions of the current problematic situation, but rather activity diagrams that represent the desired actions from various worldviews ("weltanshauung" [4]). The models become discussion aids when compared to the real world situation, and often are the source for discussion and understanding. A process of remodeling is followed to design a model that accommodates the ideas of the different world views. This should yield a model of purposeful activity that everybody can live with. When the purposeful action is taken, the situation is hopefully improved, but a further cycle of analysis is sparked. There is also a parallel process present of cultural analysis consisting of social analysis – focusing on roles, norms, and values – and political analysis – focusing on power and the commodities thereof [10].

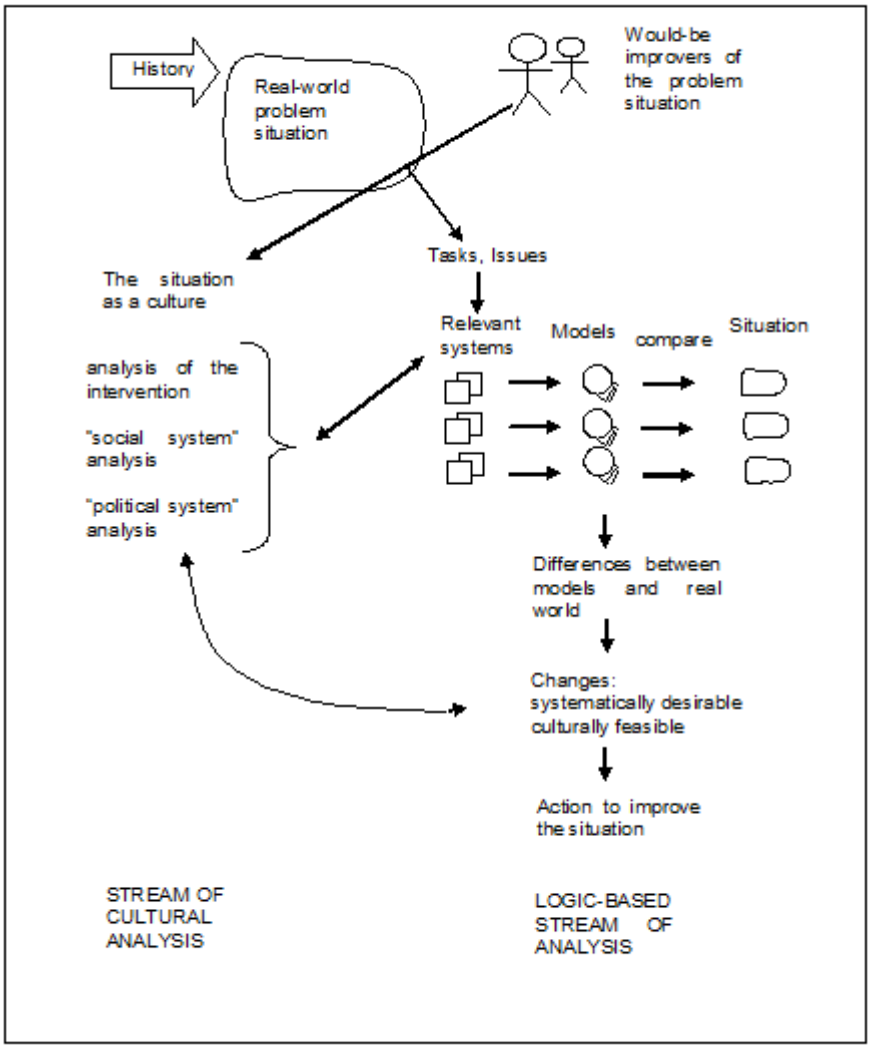

Figure 6 The overall process of the SSM [10]

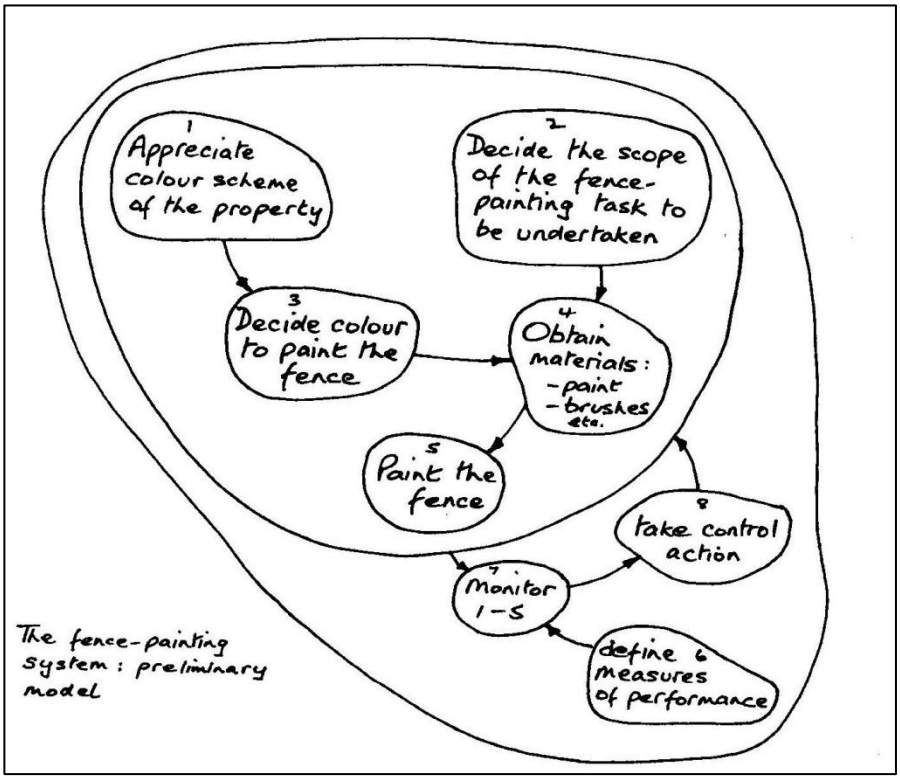

Figure 7 A simple activity diagram to paint a garden fence [9]

Model building usually starts with the identification of the transformation that is required in the problem situation. Often rich pictures are drawn to indicate the different stakeholders in the situation. A process of PQR analysis is done, where P indicates what should be done, Q how it should be done and R the higher goal to be achieved. After PQR analysis, CATWOE analysis is done to better understand the situation. The letter C depicts the customers of the action, A depicts the actors that can achieve the transformation (T). Each model is developed to represent a specific worldview (W), which is explicitly defined. The owner (O) is the party that has the power to stop the transformation. Finally E is the environmental constraints in which the transformation should take place. PQR and CATWOE lead to the development of a root definition of the purpose of the system. A good root definition includes many of the aspects of PQR and CATWOE. Next, an activity diagram is developed that demonstrates how the transformation will be done in terms of separate activities. Dependencies and flows of activities are indicated by activity numbers. Figure 7 depicts a simple activity diagram of a house owner who wants to paint his garden fence. All activity diagrams include measures for monitoring and controlling the system in terms of performance criteria. Checkland advises criteria for at least efficacy, efficiency, and effectiveness of the transformation [9].

In summary, SSM provides a set of methods to create models of the perceived action to be taken in a problem situation from different worldviews. The models generated are compared with the real problem situation and purposeful action is taken to improve the situation.

5. COMPARISON OF DATA WAREHOUSING METHODOLOGY AND THE SOFT SYSTEMS METHODOLOGY

In this section the data warehouse methodology is discussed in terms of SSM activities.

One might view the business development lifecycle shown in Figure 2 as an activity diagram for the generic action of building a data warehouse. It is an activity diagram representing the worldview of Kimball that a data warehouse can be developed from requirements collection. Already the trained SSM thinker can develop an activity diagram for developing a data warehouse from the Inmon world view. These can be compared with the actual situation in the organization. Most often a hybrid methodology is followed where some available data is investigated while the dimensional models are developed. From an SSM perspective, the model of Figure 2 should be extended to explicitly indicate performance measures of the data warehouse.

Dimensional models such as the one in Figure 4 model a perception of which data should be available in the data warehouse to satisfy user requirements. This is similar to the models in SSM which also model a conceptual understanding of the situation rather than a model of the realty. The description of modeling in data warehousing literature does not address the complexity of arriving at a specific star schema in terms of different world views or even viewpoints. It does refer to prioritization of requirement but it implicitly assumes that different modelers will come to the same star schema given the set of requirements. This assumption veers more towards the hard systems thinking approach. The simplicity of dimensional

models allows it to be used by business users which open up the possibility for it to be developed by different users to express their understanding of the requirements of the data warehouse. A process of seeking accommodation similar to that of the SSM can then be used to develop the final star schema for a business process.

6. GUIDELINES FOR USING SOFT SYSTEMS METHODOLOGY IN DATA WAREHOUSING

The section provides concrete guidelines for the use of SSM in the data warehouse development lifecycle.

1. The compelling business motivation is the transformation in SSM terms. Kimball gives a general discussion on requirements collection. This process can be enriched by applying the modeling techniques of SSM on the stated business problem.
2. The problem to be modeled using SSM centers around the key business question to answered. The use of CATWOE will highlight the importance of the owners of the operational (or legacy) systems as they have the power to stop the process: if they do not supply good quality data in a usable format the data warehouse project has to be terminated!
3. Part of model building in SSM is to set performance criteria. Often expectations of the success of data warehouses differ from user to user and from the users to the technical developers. If each interest group declares what they understand under efficacy, efficiency, and effectiveness of the data warehouse up front, many disappointments can be avoided.
4. Analysis 2 and 3 will assist in identifying the data owners and potential difficulties in the sourcing of the data. By selecting a core group of users, together with the entire data warehousing team and the owners of the source systems, all parties will understand each other's expectations and constraints in the project.
5. After the completion of the activity diagrams, each group or person representative of a specific world view should design a star schema that they believe will provide the information needed to address the business problem. These star schemas may then be compared and refined to accommodate all important requirements.
6. The project management of the data warehouse development project can viewed as a separate SSM process. As discussed in the previous section a detailed version of the lifecycle ad depicted in figure 2 can be developed. Performance measures can be incorporated to ensure that any problems are detected and addressed early. A system has an adaptive nature and ideas from agile systems development methodologies can be used to guide changes.
7. Activity diagrams may contain sub-levels; this implies that one activity can represent the outcome of an entire activity diagram. One activity in a larger diagram can be "Perform ETL" which represents an entire activity diagram.

Once users and developers are skilled in SSM many more aspects of data warehouse development will be done by using SSM activity diagrams.

7. CONCLUSIONS AND FUTURE WORK

Involvement ensures ownership and understanding. The SSM provides a set of tools to guide end-user involvement in the data warehousing project in a way that is acceptable to end-users. Most users of data warehousing projects are on managerial level in organizations. The SSM is taught in many management courses around the world as a tool for strategic management. This paper provides guidelines of extending the ideas of Kimball to incorporate the worldviews of different stake holders as suggested by SSM.

In future research a real life data warehouse project can be developed using these ideas. An action research project can then be launched to study the usability of SSM to enrich the data warehouse methodology of Kimball. This can be done by means of an action research project where a traditional data warehouse development methodology is extended to includes the ideas of the SSM as presented in this paper.

8. ACKNOWLEDGEMENT

The author wants to thank the National Research Foundation for the funding that enabled the publication of this paper.

9. REFERENCES

[1] A. Sen and A.P. Sinha, "A comparison of datawarehouse methodologies." **Communications of the ACM**, Vol. 48, no. 3, March 2005.

[2] R. Kimball, L. Reeves, M. Ross, and W. Thornthwaite, **The data warehouse lifecycle toolkit**. New York, NY: Wiley. 1998.

[3] C.W Churchman. **The systems approach**. New York, N.Y.: Delta. 1968.

[4] P. Checkland. **Systems thinking, systems practice.** Chichester: Wiley. 1981.

[5] W.H. Inmon, **Building the data warehouse.** 2nd ed. New York, NY: Wiley. 1996.

[6] F.R. McFadden, J.A. Hoffer, & M.P. Prescot, **Modern database management**. 5th ed. Reading, MA: Addison-Westley. 1999.

[7] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, & R. Becker. **The data warehouse lifecycle toolkit**. 2nd ed. New York, N.Y.: Wiley. 2008.

[8] P. Checkland & S. Holwell. **Information, systems and information systems.** Chichester: Wiley. 1998.

[9] P. Checkland & J. Poulter. **Learning for action.** Chichester: Wiley. 2006.

[10] P. Checkland & J. Scholes. **Soft systems methodology in action: Includes a 30-year retrospective.** Chichester: Wiley. 1999.