Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2010 Proceedings

Americas Conference on Information Systems (AMCIS)

8-2010

Classification of Metadata Categories in Data Warehousing - A Generic Approach

Roland Gabriel
Chair of Management Information Systems Ruhr-University of Bochum, rgabriel@winf.rub.de

Tobias Hoppe
Chair of Management Information Systems Ruhr-University of Bochum, thoppe@winf.rub.de

Alexander Pastwa
Chair of Management Information Systems Ruhr-University of Bochum, apastwa@winf.rub.de

Classification of Metadata Categories in Data Warehousing – A Generic Approach

Roland Gabriel
Chair of Management Information Systems
Ruhr-University of Bochum
44780 Bochum, Germany
rgabriel@winf.rub.de

Tobias Hoppe
Chair of Management Information Systems
Ruhr-University of Bochum
44780 Bochum, Germany
thoppe@winf.rub.de

Alexander Pastwa
Chair of Management Information Systems
Ruhr-University of Bochum
44780 Bochum, Germany
apastwa@winf.rub.de

ABSTRACT

Using appropriate metadata is a central success factor for (re)engineering and using data warehouse systems effectively and efficiently. The approach presented in this paper aims to reduce the effort in developing and operating data warehouse systems and thus to increase the ability and acceptance of a data warehouse. To achieve these objectives identifying the appropriate metadata is an important task. To avoid processing the "wrong" object data and thus compromising the acceptance of a data warehouse system, a systematic approach to categorize and to identify the appropriate metadata is essential. This paper presents such a generic approach. After investing and structuring problem situations, that can occur in data warehousing, metadata categories are identified to solve a given problem situation. A use case illustrates the approach.

Keywords

Metadata, metadata categories, data warehouse, data warehousing, design parameters.

INTRODUCTION AND RELATED WORK

Data warehouse systems (DWH systems) have become an indispensable part of the company's information logistics. They provide a cross-functional view of consolidated enterprise data (Anahory and Murray, 1997; Bucher and Dinter, 2008). The integrated processing of metadata is deemed to be a promising means of improving the use of DWH systems as well as reducing expenditure during development, operation, and maintenance of these systems (Vaduva and Vetterli, 2001; English, 1999; Ballou and Tayi, 1997). In literature many research papers focus on the integration of metadata frameworks into DWH systems (Poole, Chang, Tolbert, and Mellor, 2002; Marco and Jennings, 2004; OMG, 2003). However, they do not describe how to implement the right metadata. But some studies point out that implementing the right metadata primarily increases the acceptance of a DWH-system (e.g. Foshay, Mukherjee, and Taylor, 2007; Fisher, Chengalur-Smith, and Ballou, 2003; Watson and Haley, 1998). However, before measures for managing metadata are deployed, it is necessary to know, which metadata categories are affected in a given problem situation during data warehousing.

This paper introduces a generic approach that helps to identify metadata categories which contribute to the resolution of a given problem situation during data warehousing. Problem situations emerge when the information which is provided by a DWH cannot be processed effectively and efficiently. Initially, the metadata term is introduced (section 2). The third section focuses on the key processes, stakeholders, and problem sources of data warehousing. In the fourth section all metadata categories that are relevant for reducing a given problem situation are presented. The next two sections introduce our generic approach to identify relevant metadata categories. After a formal description of our approach in section 5 an example of how to use our approach in a real context is presented in the sixth section. Section 7 summarizes the key findings and points out future research issues.

METADATA - A DEFINITION

The widely-used definition "metadata is data about data" is seen as insufficient for a differentiated understanding of metadata (Haynes, 2004). The prefix "meta" commonly denotes data of a higher level of abstraction that describe other data of the underlying level. According to the language level theory, the data of the underlying level can also be called object data (Anandrarajan, Anandrarajan, and Srinivasan, 2004). Such data represent objects and their relationships among each other in a specific scope. Similar to other meta terms metadata always refer to an object on which statements are made. These objects are for example business processes, data structures, or rules (Marco, 2000). From the perspective of metadata these objects are object data. Hence, the meta property depends on the context it is used in.

Based on these considerations three characteristics can be identified, which help to differentiate between metadata and object data:

Object relation: Metadata always refer to corresponding object data. If a given set of metadata is not linked to object data, it cannot be called metadata.

Data description: Metadata never describe objects of reality, but only data that represent these objects.

Abstraction: Metadata are placed on a higher level of abstraction in comparison to the associated object data. This corresponds to the distinction between objects and object types or classes in the paradigm of object oriented programming.

DESIGN PARAMETERS OF DATA WAREHOUSING

The following section presents key processes, stakeholders, and problem sources which influence the development and the use of a DWH-system.

Value Chain of DWH

The purpose of data warehousing is to support decision makers and managers in their analytical tasks and the associated decision-making processes by supplying information efficiently and effectively. In this context, the DWH provides the necessary information, which is accessed by the information recipients (Anandrarajan et al., 2004). An important process for the development and the use of DWH systems is the DWH value chain (Figure 1).

During each step of DWH value chain the operational data of the source systems increase in value and are gradually condensed, until adequate information for decision support is available. The steps of the DWH value chain consist of data sourcing, data procurement, data storing, data provisioning, and data analysis.

Figure 1. Steps of the DWH Value Chain

Steps in the DWH Life Cycle

A further important process in data warehousing is the DWH life cycle. According to Kurz (Kurz, 1999) the DWH engineering process consists of five steps which are summarized in Figure 2.

The first step *Analysis and Design* comprises the creation of a specification sheet that contains the range of both function and data. Furthermore target definition, requirement engineering, detailed planning, and risk assessment are performed at this point. Based on this, the components of the DWH-system are specified. As a result several specifications are generated that define the essential requirements, e.g. regarding ETL processes, data modeling, user interfaces, security issues, and technical architectures. Based on the previously defined specifications the application is implemented and tested. Here, the individual components are developed, tested, and approved. The third step in this model is the *Use*. This step comprises any activities of the users linked to the DWH that offer support in various operational tasks. The last step in the life cycle – *Maintenance* – is the most comprehensive and longest one in addition to the step of using the DWH. Here, possible errors need to be corrected in a maintenance cycle. It also includes the DWH administration. Additionally, the system is constantly adjusted to changes in operating conditions and requirements. If so, the life cycle restarts with the step *Analysis and Design*.

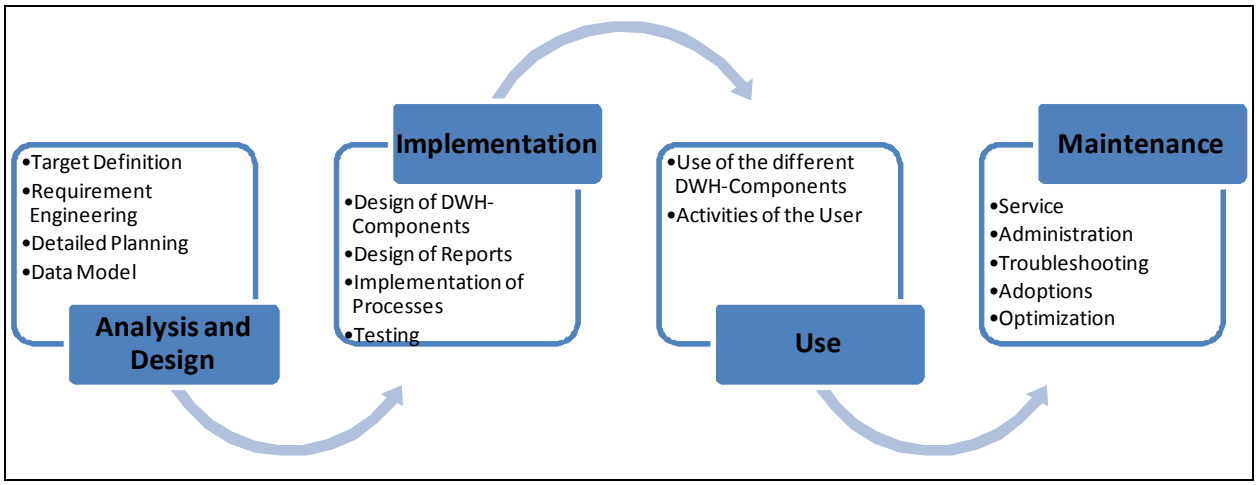

Figure 2. Steps in the DWH Life Cycle

Stakeholder

Appointing stakeholders the DWH serves to gather and provide information about who is causing or who is being affected by occurring problems. Thereby, conclusions can be drawn as to which metadata categories have to be used to solve the problem of interest from the stakeholder's point of view. Basically, people who interact with the DWH can be divided into users (customers) who demand a certain DWH product, like a report, a data mart, etc., and developers and operators (IT side) who are responsible for implementing these products and running a DWH system.

Problem Source

In case of problems in data warehousing the causes of the problems need to be analyzed. Problems mainly derive from two areas: On the one hand there are problems that solely occur because of technical reasons. These are network errors, server outages, unavailable DWH components, as well as failed accesses to altered or faulty data, for instance. Also reports that cannot be executed due to software errors or limited access privileges belong to this technical problem area.

On the other hand there are functional-related problems. Such problems occur if essential business management knowledge about the data in the DWH is missing. Often such problem situations occur especially in connection with the use of DWH systems. When knowledge or understanding about the meaning, origin, structure, and quality of the data is not available in sufficient quantities for stakeholders, tasks can only be performed in a limited way.

METADATA CATEGORIES

This section focuses on all metadata categories that are appropriate to serve as a starting point for countermeasures for mitigation and elimination of problem situations in data warehousing. While many authors distinguish between functional and technical metadata and evaluate them differently (see e.g. Inmon, 2001; Devlin, 1997), Auth (Auth, 2004) identifies metadata categories which are able to describe object data in a DWH in a purpose-oriented and systematic way. The classification presented in this paper refers to the proposition of Auth (Auth, 2004).

Category 1: Terminology

In order to integrate operational systems it is necessary to identify the data objects, which are stored in the DWH, and to standardize their labeling and meaning. Furthermore, homonyms and synonyms need to be identified as well as false and vague terms, which are used for describing the objects, have to be corrected (Winter and Strauch, 2004). Herein is the role of metadata which are responsible for the terminology of the used dimensions, dimension elements, dimension hierarchies, and measures.

Category 2: Data Analysis

Metadata of this category provide an overview of the processing and usage of object data (Jarke, Lenzerini, Vassiliou, and Vassiliadis, 2000). For the purpose of data analysis Online Analytical Processing (OLAP) and/or data mining methods have to be considered. Based on the central DWH data basis, data are often transferred into data marts that are optimized for analytical purposes.

Category 3: Organization Reference

The interrelationship between data and the associated business organization is given on the one hand by the tasks that require the data for execution (process organization), and on the other hand by the stakeholders, who access these data to fulfill their tasks (organizational structure) (Winter et al., 2004). Metadata of the category organization reference describe where object data are created or imported, and how they are used in business processes. In addition, metadata of this category document access privileges as well as privacy classifications (Devlin, 1997).

Category 4: Data Quality

Furthermore, metadata provide information on the quality of the object data. Thereby, the quality of data can generally be evaluated by two measures: The objective correctness of data values (e.g. precision, consistency) and the subjective aptitude of data to satisfy a certain information need.

Category 5: Data Structure and Data Meaning

The fifth metadata category describes the structures of data as well as their meaning. During the process of extracting, transforming, and loading (ETL) data from the source systems, data objects are grouped to data object types according to their relationships among each other. Combined data object types are called data structures (Laudon and Laudon, 2005).

Category 6: System Reference

Metadata of this category inform about specific characteristics of a DWH architecture (Hufford, 1997). If a user wants to locate data it is helpful to know which component of a DWH system manages these data (e.g. a particular data mart). Besides the software components, information about the used computers, networks, and storage systems, on which the software components run, are recorded as metadata. Further relevant metadata are information about locations and organizational responsibilities for software and hardware (Devlin, 1997).

Category 7: Data Transformation

Metadata belonging to the category data transformation describe the path object data takes from the source systems through the layers of the DWH to the analytical applications (Jarke et al., 2000). For this purpose single transformation steps are combined to transformation processes. Such descriptions of data transformations are especially used by experienced users who interpret the results of analyses. Hence, the acceptance of these results will increase.

Category 8: Metadata History

When data changes over time the changes have to be recorded with the help of a versioning system. Such a system ensures that data objects are not overwritten. Instead, during every data changes a new version of the data set is created (Tannenbaum, 1994). Recording such changes is particularly important for time series analyses. A complete object data history can only be achieved by creating a metadata history simultaneously.

FORMAL DESCRIPTION OF THE GENERIC APPROACH

The following section presents a formal description of our generic approach. The approach consists of six steps helping to determine metadata categories which are appropriate to solve a given problem situation in data warehousing.

1st Step: Identifying and Classifying Sub-Problems

Initially a given problem situation which is the starting point for applying our approach has to be analyzed and divided into sub-problems. Subsequently, it has to be determined, which of the sub-problems affect the design parameters of data warehousing described in the third section. Both the design parameters ("Value Chain" (*V*), "Life Cycle" (*L*), "Stakeholder" (*S*), and "Problem Source" (*P*)) including their specifications and the sub-problems (*N*) can be represented as:

$$V = \{v_1, v_2, \ldots, v_5\} \rightarrow |v| = 5$$
$$L = \{l_1, l_2, \ldots, l_4\} \rightarrow |l| = 4$$
$$S = \{s_1, s_2\} \rightarrow |s| = 2$$
$$P = \{p_1, p_2\} \rightarrow |p| = 2$$
$$N = \{1, \ldots, n\}$$

The resulting matrix *D* stands for a set of characteristics of design parameters according to the sub-problems, where *r* is a set of characteristics which is affected by one sub-problem:

$$D = V \times L \times S \times P$$
$$r \in D \; with \;\; r = (v, l, s, p)$$

The following expression guarantees that each sub-problem affects only one set of characteristics:

$$\forall\, i \in N \; \exists\, r \in R : c(i) = r$$

2nd Step: Identifying Groups of Similar Problems

After identifying possible sub-problems, similar sub-problems have to be grouped. Each group *g* consists of various sub-problems which affect identical combinations of design parameters of data warehousing. The formal description is:

$$G = \{g_1, \ldots, g_k\} \; with \; k \leq |N|$$
$$and \; it \; is:$$
$$g_i \cap g_j \varnothing \; \forall\, i \neq j \quad \text{(The groups are pairwise disjoint.)}$$
$$\bigcup_{i=1}^{k} g_i = N \quad \text{(Each sub-problem is in a group.)}$$
$$c(i) = c(j) \; \forall\, i, j \in g_l \; \forall\, l \quad \text{(In one group, the characteristics of all sub-problems are equal.)}$$
$$c(i) \neq c(j) \; \forall\, i \in g_l,\; j \in g_m \; with \; m \neq l \quad \text{(In different groups, the characteristics of the sub-problems are different.)}$$

3rd Step: Weighting the Sub-Problem Groups

The next step aims to classify the sub-problem groups. For this purpose the sub-problem groups have to be weighted. The weighting factors of each group are determined by the number of contained sub-problems. In order to weight the sub-problem groups the weighting factor (*w*) of each group are determined by the number of contained sub-problems:

$$w(g_l) = |g_l| \; \forall\, 1 \leq l \leq k$$
$$c(g_l) := c(i) \; with \; i \in g_l$$

4th Step: Assigning Metadata Categories to the Design Parameters

The next step is to ascertain for a given problem situation, which metadata categories influence which design parameters of data warehousing. The result is a Cartesian product of the power set ($\wp$) of the design parameters and their influencing metadata categories:

$$X \leq \wp(V) \times \wp(L) \times \wp(S) \times \wp(P)$$
$$|X| = 8$$

The weighting factors are transferred to the corresponding metadata categories for all sub-problem groups. Subsequently, the sum of all weighting factors is calculated for each metadata category:

$$\forall g_1 \ldots g_k : Y_x = \sum_{i=1}^{k} f_x(g_i)$$

$$f_x(g_i) = w(g_i) \text{ if } v \in \upsilon, l \in \lambda, s \in \sigma, p \in \pi \text{ with } c(g_l) = (v, l, s, p)$$

and

$$x = (\upsilon, \lambda, \sigma, \pi) \in X \text{ with } \upsilon \leq V, \lambda \leq L, \sigma \leq S, \pi \leq P$$

and else

$$f_x(g_i) = 0$$

5th Step: Sorting Metadata Categories

After assigning metadata categories to the design parameters of data warehousing the metadata categories have to be sorted according to their importance as for solving a given problem situation. To determine the priority (Π) the sums (Y_x) are sorted in descending order:

$$\Pi = desc(Y_x)$$

The higher the priority, the greater is the potential of the metadata category to contribute to the problem solution.

6th Step: Interpretation of the Results

Finally the results have to be interpreted in order to identify metadata categories which are appropriate for solving a given problem situation in data warehousing.

EXEMPLIFICATION OF THE PROCEDURE

The following section exemplifies the classification of metadata categories to solve problem situations in data warehousing. In order to make our approach clear we will look closer at a typical problem we have observed in sales reporting many times. The use case is as follows:

Every first Wednesday of the month a sales meeting is held. Besides other participants, this meeting is attended by the division manager and the responsible division marketing manager. The meeting is held in order to discuss the results of the last month. In particular, the sales figures of the different sales teams and branches are critically assessed and compared to the forecasts.

The information basis of the division manager is a variety of reports which are produced and provided by the controlling department. These reports contain information which is provided by a DWH system. The DWH system consolidates data from different source systems. One of the source systems is a customer relationship management system. This system provides the sales figures which are used for decision making by the division marketing manager. That is, the division marketing manager accesses information from the customer relationship management system directly without using the DWH.

In the course of the sales meeting, the participants find out that the sales figures in the reports of the division manager and the sales figures in the reports of the division marketing manager differ from each other. After analyzing the reasons for this inconsistency following sub-problems have been identified.

Reasons for inconsistent sales figures

Insufficient Measure Labeling in Reports

The measure “sales” did not indicate whether these sales were net sales or gross sales.

Non-Uniform Measure Definition

Another cause of error was the ambiguous definition of measures. In the use case described above it was unclear, whether any discounts and bonuses were subtracted in the reports, thus reducing the revenue.

Rounding Differences

Since the sales figures were rounded to one decimal place in the reports of the division marketing manager, inconsistencies in subsequent measures that were calculated on the basis of these rounded values were the consequence. Especially in case of aggregating data series these rounding differences can have an enormous impact on the results.

Miscalculation of Measures

Besides ambiguous definitions of measures, errors can also arise in the mapping of the transformation logic. If DWH developers, as it happened in our case, implement incorrect calculation rules, the sales figures will be faulty.

Coexistence of Sales Reports

The longer a DWH is used in a company, the more calculation rules for measures are supposed to be altered (e.g. due to legal requirements). In order to be able to analyze older data (if necessary), reports containing the old calculation logic remain. This was the case in the report of the division marketing manager. In addition new reports were created which base on the new calculation logic. Hence, old and new reports coexisted.

Report Creation at Different Times

The significance of a report depends on its creation date. Since one and the same report was created at different dates, the results were not identical due to the time lag in between.

Incomplete Data or Values in the Reports

ETL processes are executed at different times in order to fill the DWH with up to date data which actually are needed by the users. This might be once a day, once a week, or once a month. Since in our use case an ETL process has terminated in error and has not finished at the time of reporting respectively the report data which were used by the division manager were not complete.

Steps for the Classification of Relevant Metadata Categories in Data Warehousing

The following six steps help to determine the metadata categories which need to be implemented to obtain a solution for a given problem case.

1st Step: Identifying and Classifying Sub-Problems

In the use case described above it is evident that labeling measures affect the quality of the reports, for instance. In this context it has to be pointed out that a sub-problem can affect different design parameters and therefore must be split into different sub-problems. This is the case in the "Measure Definition" (sub-problem 2a to 2c) and "Rounding Differences" (sub-problem 3a to 3b). It is also evident that an insufficient measure definition is problematic during the implementation or development of a report (sub-problem 2b to 2c). The affected steps are data procurement and data storage. Considerations are to be made in an analogous manner for the other sub-problems. Table 1 depicts the results for the present use case.

2nd Step: Identifying Groups of Similar Problems

As seen in Table 1 sub-problems 1, 2a, 3a, and 5 can be merged to one group, since they all share the characteristics "Data Analysis", "Use", "User", and "Functional" (cf. color highlighting in Table 1). Altogether, five different groups can be identified.

3rd Step: Weighting the Sub-Problem Groups

Table 1 indicates that group 1 is composed of four sub-problems (cf. color highlighting in Table 1). Hence, it receives the weighting factor of 4. For the second and fifth group the factor amounts to 2, for group three and four it is 1. Alternatively, in this step also non-linear weighting procedures can be used.

Design Parameter	Value Chain					Life Cycle				Stake-holder		Problem Source		
Characteristic / Sub-problem	Data Source	Data Procurement	Data Storage	Data Provision	Data Analysis	Design	Implementation	Use	Maintenance	User	Developer / Operator	Functional	Technical	Group
1. Measure Labeling					x			x		x		x		1
2a. Measure Definition					x			x		x		x		1
2b. Measure Definition				x			x				x	x		2
2c. Measure Definition		x					x				x	x		3
3a. Rounding Differences					x			x		x		x		1
3b. Rounding Differences				x			x				x		x	4
4. Calculation				x			x				x	x		2
5. Several Reports					x			x		x		x		1
6. Date of Retrieval					x			x		x			x	5
7. Incomplete Data					x			x		x			x	5

Table 1. Classification and Grouping of Sub-Problems

4th Step: Assigning Metadata Categories to the Design Parameters

This step aims to assign metadata categories to the design parameters of data warehousing. The result of this step is a "mapping table", as is shown in Table 2. The assignments can be the agreed-on result of a workshop which is attended by all users, developers, and operators of the DWH system in which a given problem situation occurs.

Design Parameter	Value Chain					Life Cycle				Stake-holder		Problem Source	
Characteristic / Category	Data Source	Data Procurement	Data Storage	Data Provision	Data Analysis	Design	Implementation	Use	Maintenance	User	Developer / Operator	Functional	Technical
Terminology		x	x	x	x		x	x		x	x	x	
Analysis					x			x		x		x	
Organization Reference	x				x			x	x	x	x	x	x
Quality					x		x	x		x		x	x
Structure / Meaning	x	x	x	x	x	x	x		x	x	x	x	x
System Reference	x		x	x		x	x		x		x		x
Transformation		x		x	x		x	x	x	x	x	x	x
History of Metadata			x		x		x	x		x	x	x	x

Table 2. Example of a „Mapping Table"

The mapping table which is depicted in Table 2 indicates that a consistent terminology of dimensions and measures is an important precondition for the successful analysis of sales data (see the blue mark) in the past. To use information about the preparation and use of object data during the process of data analysis, metadata of the category "Data Analysis" need to be integrated. According to our experience in sales reporting information about the development and use of sales data were not or only insufficiently available. Consequently, metadata relating to the organization offer potential benefits during data analysis. As for the sales reporting both the objective correctness of the data values as well as subjective suitability of the sales data were classified as critical, future data analysis need to include metadata that inform about the quality of the object data. Further, it should be noted that in the data analysis of sales reporting a detailed knowledge of the reported data objects and data structures is required, so that metadata of the category "Data Structure and Data Meaning" must be implemented. Furthermore, to analyze sales data in context of the value-added process of the DWH, it must be ensured that there is a corresponding metadata history, and all performed transformation steps prior to the measure creation are considered.

As for the process of data analysis, the same procedure is applied for the other design parameters. That is, each purpose the DWH is used for and each metadata category must be reviewed against the backdrop of the question how it has affected the design parameters of data warehousing in the past.

5th Step: Sorting Metadata Categories

The priorities which result from the sorting process can be found in the lower part of Table 3. To determine the priority the information from the top of the table must be considered. The table shows that the combination of the design parameters "Data Analysis", "Use", "User", and "Functional Background" affect all of the metadata categories except "Structure / Meaning" and "Reference to System". To determine this result, an AND Operator is used to combine the characteristics of the design parameters. Consequently, all metadata categories of sub-problem group 1 receive a weighting factor of 4. The proceedings for the other sub-problem groups are analogous. Afterwards for each metadata category a total value is determined that results from the row totals of the individual weighting values. The highest value receives the first priority. The first two priorities have a colored background in Table 3. The company the scenario refers to can reduce or at best avoid the occurrence of inconsistent measure values completely in the future, if metadata of the categories "Transformation" and "Terminology" are used in the DWH. The other identified categories with lower priorities can obviously also be used in solving problems, but are not as beneficial as the first two priorities.

Group	**1**	**2**	**3**	**4**	**5**		
Weighting Factor	4	2	1	1	2		
Value Chain							
Data Source							
Data Procurement			x				
Data Storage							
Data Provision		x		x			
Data Analysis	x				x		
Life Cycle							
Design							
Implementation		x	x	x			
Use	x				x		
Maintenance							
Stakeholder							
User	x				x		
Developer		x	x	x			
Problem Source							
Functional	x	x	x				
Technical				x	x		
Categories of Metadata						**Σ**	**Priority**
Terminology	**4**	**2**	**1**	**1**	**-**	**8**	**2**
Analysis	**4**	**-**	**-**	**-**	**-**	**4**	**4**
Organization Reference	**4**	**-**	**-**	**-**	**2**	**6**	**3**
Quality	**4**	**-**	**-**	**-**	**2**	**6**	**3**
Structure / Meaning	**-**	**2**	**1**	**1**	**-**	**4**	**4**
System Reference	**-**	**-**	**-**	**-**	**-**	**-**	**-**
Transformation	**4**	**2**	**1**	**1**	**2**	**10**	**1**
History of Metadata	**4**	**-**	**-**	**-**	**2**	**6**	**3**

Table 3. Results

6th Step: Interpretation of the Results

The metadata category "Data Transformation" primarily supports the developers. When transformation steps are well documented and described, calculations of measures can be understood faster and possible errors located easier. Documenting transformation processes prior to their implementation significantly contributes to more efficient data integration. Prepared reports on the progress of the process execution ensure that. For example, at the time of reporting all necessary amounts of data are loaded into the DWH. Stored metadata help users understand the data sources in particular. The resulting level of transparency improves user acceptance and facilitates data interpretation.

In order to standardize data objects which are processed in a DWH metadata of the category "Terminology" are also required. The systematic collection of terms in form of a glossary serves a continuous documentation. Calculation schemes or even formulas can, for example, substantiate the mere definition of a measure. Synonyms and homonyms are also important components of such a catalog. The homogenization of reporting results is a further beneficial aspect. Users are able to exploit

and interpret contents of reports more quickly. Misunderstandings about the use of terms are minimized. During the process of standardization and homogenization all stakeholders need to be included in order not to compromise the acceptance of a unified system.

SUMMARY AND FURTHER RESEARCH

This paper presents a generic approach that allows identifying problem-specific metadata categories. The approach was presented by means of a use case that is often found in sales reporting. The starting point were inconsistent measures in sales reports.

Our approach provides value, when applied in a feasibility study, and thus in an early phase of (re)engineering a DWH. The identified metadata categories help to gain an early impression of which goal will be achieved and what personnel, technical, and organizational capacities are needed to resolve the identified problem. This has in turn further implications for project planning.

Further on, the approach offers the advantage to consider lessons learned from past projects and applications that relate to a given problem situation. In our approach these lessons learnt are contained in the mapping table (Table 2). In order to create such a mapping table we advice to hold a workshop which is attended by the DWH users, developers, and operators. Hence, a certain amount of preparatory work must be done.

Furthermore, our approach provides the flexibility to integrate additional design parameters of data warehousing which have not been mentioned in section 3. The presented design parameters reflect only a part of the reality in DWH. Since this paper focuses on the generic approach for the classification of metadata categories, a most complete and detailed thematization of all design parameters was waived. Nevertheless, it has to be examined on how further design parameters complement our approach in a useful way and help to improve identifying problem-specific metadata categories in a given context. However, the number of design parameters does not affect the generic approach this paper has focused on.

REFERENCES

1. Anahory, S. and Murray, D. (1997) Data Warehousing in the Real World: A practical guide for building Decision Support Systems, Addison-Wesley, Harlow.
2. Anandrarajan, M., Anandrarajan, A., and Srinivasan, C. R. (2004) Business Intelligence Techniques, Springer, Berlin.
3. Auth, G. (2004) Prozessorientierte Organisation des Metadatenmanagements für Data-Warehouse-Systeme, Books on Demand, Norderstedt.
4. Ballou, D. P. and Tayi, G. K. (1997) Enhancing Data Quality in Data Warehouse Environments, *Communications of the ACM*, 42, 1, 73-78.
5. Bucher, T. and Dinter, D. (2008) Process Orientation of Information Logistics - An Empirical Analysis to Assess Benefits, Design Factors, and Realization Approaches, in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, Waikoloa, IEEE, 392.
6. Devlin, B. (1997) Data Warehouse: from architecture to implementation, Addison-Wesley, Reading.
7. English, L. P. (1999) Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits, Wiley, New York.
8. Fisher, C. W., Chengalur-Smith, I., and Ballou, D. P. (2003) The Impact of Experience and Time on the Use of Data Quality Information in Decision Making, *Information Systems Research*, 14, 2, 170-188.
9. Foshay, N., Mukherjee, A., and Taylor, A. (2007) Does data warehouse end-user metadata add value? *Communications of the ACM,* 50, 11, 70-77.
10. Haynes, D. (2004) Metadata for Information management and retrieval, Facet, London.
11. Hufford, D. (1997) Metadata repositories: the key to unlocking information in Data Warehouses, in Ramon C. Barquin and Herbert A. Edelstein (Eds.) *Planning and designing the Data Warehouse*, Prentice Hall, Upper Saddle River, 225-262.
12. Inmon, W. H. (2001) An Illustrated Taxonomy of Metadata, White Paper, l.c.
13. Jarke, M., Lenzerini, M., Vassiliou, Y., and Vassiliadis, P. (2000) Fundamentals of Data Warehouses, Springer, Berlin.
14. Kurz, A. (1999) Data Warehousing, enabling technology, Mitp, Bonn.

15. Laudon, K. C. and Laudon, J. P. (2005) Management Information Systems, Managing the Digital Firm, Prentice Hall, Upper Saddle River.
16. Marco, D. (2000) Building and managing the Meta Data Repository, a full lifecycle guide, Wiley, New York.
17. Marco, D. and Jennings, M. (2004) Universal Meta Data Models, Wiley, New York.
18. OMG (2003) Warehouse Metamodel (CWM) Specification, Version 1.1, Volume 1, OMG, l.c.
19. Poole, J., Chang, D., Tolbert, D., and Mellor, D. (2002) Common warehouse metamodel: An Introduction to the Standard for Data Warehouse Integration, Wiley, New York.
20. Tannenbaum, A. (1994) Implementing a Corporate Repository, the model meets reality, Wiley, New York.
21. Vaduva, A. and Vetterli, T. (2001) Metadata Management for Data Warehousing: An Overview, International Journal of Cooperative Information Systems, 10, 3, 273-298.
22. Watson, H. J. and Haley, B. J. (1998) Managerial Considerations*, Communications of the ACM*, 41, 9, 32-37.
23. Winter, R. and Strauch, B. (2004) Information Requirements Engineering for Data Warehouse Systems, in Hisham M. Haddad (Ed.) *Proceedings of the 2004 ACM Symposium on Applied Computing*, Nicosia. ACM, 1359-1365.