

Exploiting DWH Metadata for Knowledge Management

Berthold Weismann, BADENIA Bausparkasse AG
Wilfried Schollenberger, WS Unternehmensberatung und Controlling-Systeme GmbH

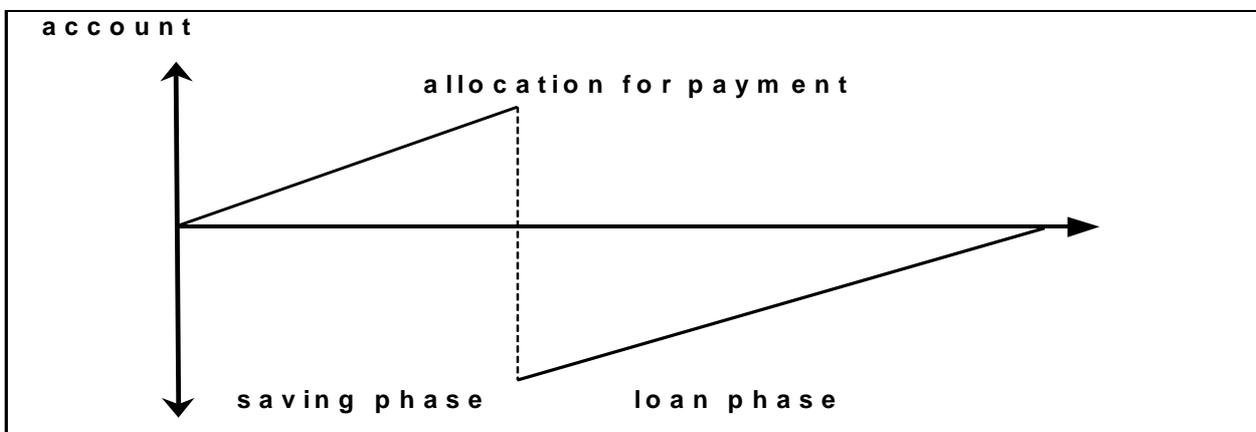
Abstract:

The BADENIA Bausparkasse AG attaches great value to a proper modelling and documentation of their data warehouse. The information must be provided in a manner that other developers, knowledge workers, and new employees can easily recognize what exactly is in the data warehouse and are able to use this data for their work. This task must be considered in the context of knowledge management. Therefore, the BADENIA Bausparkasse uses the tool XDWH of the WS Unternehmensberatung.

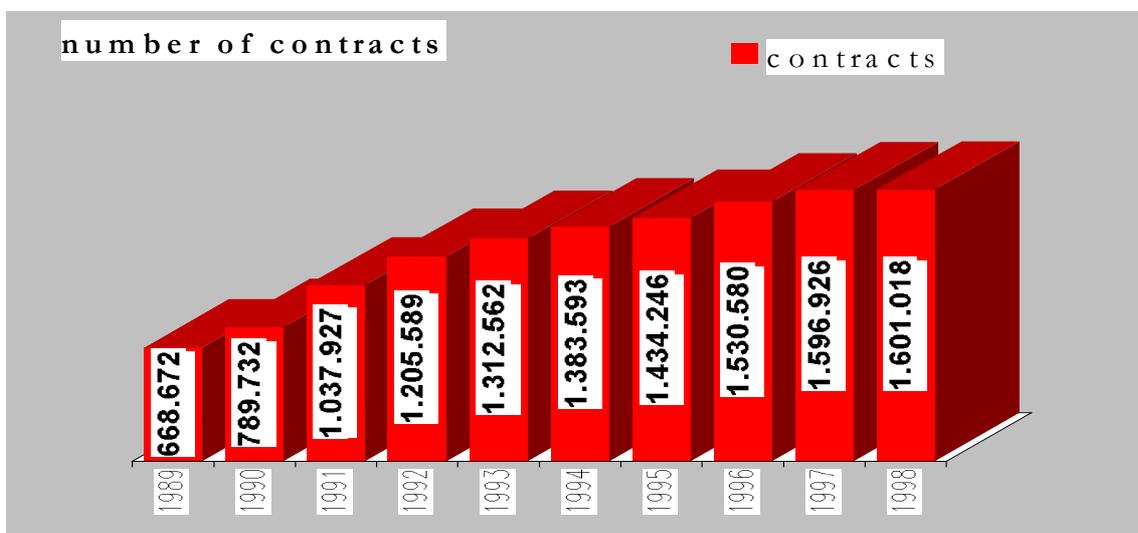
The paper discusses the necessary contents of a metabase and an approach to publish this information via the Intranet. It is shown how to integrate the creation and maintenance of this meta information into the development process in a way that it is not simply additional work. The metabase must support the work of the developers in a way that it is expectable that the metabase will always reflect the current state of the data warehouse.

1. The BADENIA Enterprise Planning and Controlling Department (UPC)

The Badenia Bausparkasse is a building society. A German building society is a credit institution for the provision of housing, capital formation and personal old age provision with a very special system and structure. This special system, called the allocation system, comprises three phases: the savings phase, the allocation for payment phase and the loan phase. During the savings phase the investor makes regular deposits. Together with the combined deposits of other investors and the amortization payments of the customers during phase three this pool of inflowing money is used to pay out the loans. The major advantage is obvious: the allocation system provides loans at fixed low interest rates independent from the capital market. So one of the major tasks facing the managers of a building society is to maintain a steady flow of new deposits.



After an enormous growth in the last years the Badenia Bausparkasse is now the number four of the private building societies in Germany. With a balance-sheet total of 4,5 billion euro, almost 1,5 million customers with 1,6 million contracts and an inflow of savings of 880 million euro. The Badenia Bausparkasse is part of the Aachener und Münchener Group.



The BADENIA sales force is very heterogeneous. To report and control them is one big challenge to the UPC (Unternehmensplanung und Controlling) the BADENIA enterprise planing and controlling department. Other challenges in this department are:

- to develop new products
- management accounting
- to develop and maintain a reporting system
- to monitor the quality of the sales force

The UPC is using the SAS software for sales controlling since 1993. An EIS (presented on SEUGI 1993) and a batch reporting using the same database were developed. Up to now other applications developed with the SAS software are added:

- quality controlling
- a model that simulates the in- and outflow of deposits and funds and which allows a forecast on revenues and expenditures (LIRE); the model was developed with a tool from WS Unternehmensberatung
- special reports to meet the requirements of the German Banking Law
- applications to analyse our contracts and investors

The UPC is the main provider for information in the company because of:

- the information they can derive from data
- the databases they developed
- their ability to create ad hoc queries.

2. Our Data Warehouse (DWH) becomes a Foundation for the Evolution of UPC in the Future

The development of the various applications was done by the UPC itself with few external consulting. The advantages of this strategy are the following:

- the knowledge workers keep control of their own applications
- they have learned a lot about the data and the data quality
- they have built up skills and knowledge

Up to now the data for applications have been loaded directly from the operational data source (ADABAS). But now the realisation of new applications (e.g. loan-reporting, MIS for our sales forces, risk controlling) will benefit from a single enterprise Data Warehouse (DWH).

To the UPC having an enterprise DWH means:

- having the data available for all kind of analysis and reports
- having historical data available so that we can easily reproduce a former state of data and create time series.
- having everything documented in an integrated and consistent system, so everyone can have access to it
- the department grew so co-ordination became more necessary; we want to manage our joint knowledge and provide it in a written and easy accessible manner
- a quick response to questions
- a well-documented database that is easy to understand

3. Technical Approach

The BADENIA DWH contains two levels. An enterprise DWH called level 1. This level will provide the data for five existing and 3 new realised applications. The application layer called level 2 contains data derived from level 1. These data are specially prepared for the different applications.

3.1. In Level 1 a Normalized Relational Approach is used

UPC puts great emphasis to the fact that the relational model describes all the parts of interest of the company such as structures or relations and creates an image of what can be observed in the company. In other words it describes the context of the DWH completely and it is not necessary to look at the data itself in order to know what is in a table.

The physical database is implemented as SAS datasets and matches the ER-modell one-to-one. This gives us the security that every value is loaded once and only once. In other words there is no loss of information by loading data, and the maximum flexibility for reports and analyses that may come in the future is preserved.

Storing historical information in the DWH enables UPC to generate every historic state. There are three different kinds to manage history:

- the date with the specific attributes is active in an interval; interesting for data with less change over time
- the date is read at defined moments; this kind of history is used to handle data often changed
- date is aggregated in defined intervals

3.2. In Level 2 Denormalized Datasets are used

Level 2 contains denormalized SAS datasets that serve as data source for applications. Depending on the application they are used directly or as an interface file that is further processed by the application.

Again it is very important that it is documented how these datasets are derived from level 1. Everyone must be able to trace back the whole process to the original source in the operational data.

4. The Role of Data Modelling

Based on an experience of 7 years with the SAS software UPC already had confidence to the SAS data management features and knew how to exploit them. So the main tasks in data modelling were

- to get a profound understandig of the company
- to understand how the company is depicted in the operational data
- to ensure that level 1 of our DWH contains all relevant data

Therefore, UPC decided to integrate technical and special information in one modelling and documenting system. Therefore they selected XDWH from WS.

For a further improvement of the quality of work a 3-step-approach is used:

1. UPC makes a detailed list of requirements to be reviewed by external developer.
2. The developer enters the information into the modelling tool XDWH. This tool automatically transforms the information into linked HTML documents.
3. UPC reviews for the correctness and completeness of the information.

With this approach the developer and all members of the UPC get the same understanding and a common language. Further it is ensured that the documentation is complete and fully matches the contents of the DWH.

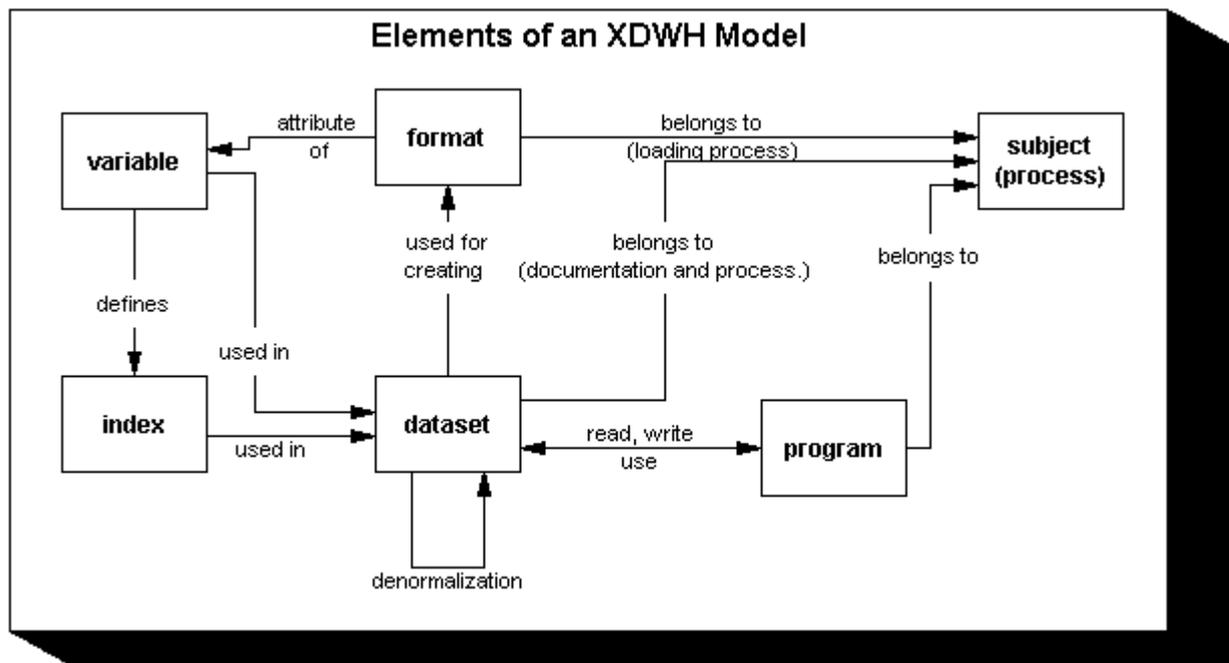
5. The XDWH Tool for Enterprise Data Warehouses

We are now going to explain XDWH and the methods behind it in more detail.

5.1 Background

The XDWH tool was further developed from the cross reference tools which were introduced at SEUGI 1995 in Stockholm ("Let's talk about Case"). The main reason was to support an efficient method for modelling an enterprise DWH which is implemented using the SAS System for storing the data. In principle it is a normalized relational model. The information is defined as attributes and allocated to the objects in the business. From these objects, e.g. contracts, sales representatives, customers, the entities are directly derived. It is very important that the model describes entities which knowledge workers recognize in the company, and that the physical implementation matches this model one-to-one. In this way the greatest flexibility for reporting is preserved, and it is easy to recognize how the data for a new report or evaluation can be extracted. Describing the content, the model must include user-defined SAS Formats and Informatms to identify the coding schemes. They can also be used to identify the type of data, like currency and rates. Furthermore SAS Formats can be used to provide consistent groupings, e.g. for age and terms.

In order to support consistency, the variables and indexes are defined once in global dictionaries before they are used in tables. So variables like age, sex, and city will always have the same type, label, and format in every dataset. Their definitions are only maintained once.



XDWH includes the special aspects of each element together with all technical aspects in one consistent system. Precise definitions in the modelling phase help to detect small but important differences in the views of different departments. Therefore, the special and the physical information about an element, e.g. a variable or a dataset, are to be entered and reported at the same time.

This information should be accessible without using any specialized software. Furthermore it should be current, since the DWH may be quite often enhanced or modified, e.g. after reorganizations in the company.

Therefore, we provide the complete documentation in HTML format with a lot of links. Most of this documentation is created automatically from the model database.

5.2 The Scope of XDWH

In order to provide full information for the special departments, XDWH covers:

- the enterprise DWH to show what is available,
- the loading processes to show where the data comes from,
- the denormalized datasets which are created for applications (Data Marts).

On every level it is possible to record how problems, errors, and exceptions are handled. This is done by entering remarks that can be attached to any number of variables, datasets and programs. E.g. we record the methods for error checking and the way to handle errors. These notes are linked to the involved programs, datasets, and variables. Later, when the variable is used in a Data Mart, the user will automatically get the links to the relevant methods for error checking, when he loads the documentation of the Data Mart.

Though code generation is not a mature issue, it improves the usage of the tool when XDWH creates the code for all empty datasets and for the creation of user defined SAS Formats and Informats from datasets via the CNTLIN-option of PROC FORMAT. So it really saves work, when a developer maintains the documentation first.

5.3 Developing and Documenting the Enterprise Data Warehouse

Usually the development of the enterprise DWH model is broken down into subjects as areas of interest. The first step is to write down the special requirements in an HTML document. All requirements have the same standardized format with the following topics:

- Set: What are the objects in the enterprise? Example: "Contracts with customers, their balances and turnover".
- Handling: What has to be observed when collecting and storing the data? Special remarks when the handling is not obvious, e.g. how are securities allocated to loans?
- Data source: Where does the data come from? Operational databases and applications.
- Periods: How long shall the data be kept? Which periods shall be available? Frequencies of updates.
- Relations to other subjects: Relations where one entity is covered by another subject.
- Future extensions: Which extensions are planned or can be foreseen for the future?

The next step is the identification of the entities and their relations. An Entity-Relationship-diagram (ER-diagram) is a good visual mean for that purpose. It can be created by any good charting tool which is able to create GIF-files. The ER-diagram for subjects is refined until each entity and each many-to-many relation has exactly one corresponding table in the physical implementation. Using techniques like specialization and splitting of entities according to the type of handling history leads to different entities in the model. So everybody gets a good visual impression of how to merge the data. Additionally the mayor entities for all subjects and their relations are shown in an overview diagram. XDWH creates an HTML file for each subject which includes the corresponding GIF-file.

All other information about the contents of the DWH is entered via XDWH and stored in a SAS database.

The clear and comprehensive definition of each dataset and variable is the most important source for the HTML documentation and the foundation for the further development of the loading procedures. So specialist workers can check what is in the DWH and whether the definitions are correct. Developers take that information to create the loading processes.

5.3.1 Subjects and logical Views

To improve the overview, many tables will be allocated to more than one subject. A subject in this sense is an area of interest and the number of subjects which are defined depends on the needs of the different users. Subjects are used to answer to questions like "What do we know about sales?".

To tell the users what information about an entity can be derived from the DWH, denormalized logical views are defined. In these views the different tables are linked to a "leading" table and the HTML documentation will have the links to the appropriate tables in the model. These views are used to answer to questions like "Which variables can be allocated to a sales representative?" and "Which variables are allocated to a loan?".

5.3.2 Dealing with History

Handling the aspects of history while creating the model is not appropriate. This would lead to one-to-many and many-to-many relations where we would expect one-to-one. Instead the model and the Entity-Relationship-diagrams are set up for one single point in history. The type of handling history becomes an attribute of each table. It is taken into account, when logical views and Data Marts are defined. Depending on this type the appropriate variables for handling history are automatically included in the generated code for creating empty datasets.

5.4 Data Marts

It is good practice to define and create the basic datasets that are used by applications as Data Marts. These datasets may be processed further by the application, e.g. create an MDDB for SAS/EIS. XDWH handles these Data Marts like subjects and the datasets similarly to logical views. In this way the developers and the users of an application can check the data they get from the DWH and receive all information about it that is stored in the enterprise model.

5.5 Documenting the Programs

Everybody who gets data from the DWH must be able to reconstruct the processing that delivered the data to him. This gives confidence and enables him to check the processing back to the DWH tables and the original data, when he finds some anomalies which he does not trust.

Therefore, every program is recorded with a description and the datasets it reads and writes, including temporary, error and protocol datasets. The documentation of a dataset will contain links to all programs that access. Additional subjects can be created to describe the data flow, the programs, and error checking. Again the flow charts are created outside of XDWH using the charting tool.

Experiences

For us, UPC at BADENIA Bausparkasse AG, it is very important that we achieved a number of important goals:

1. With the help of WS Unternehmensberatung we implemented an efficient method to work with external developers. With our 3-step-approach the quality of our work and results was greatly improved.
2. With this method of modelling we early became aware of the special details and could handle them in a reasonable manner and in short time.
3. As a result of the documentation we developed a common naming and a common understanding of the whole enterprise and of details.
4. The easy access of the documentation supports our daily work.

In the beginning it was quite difficult to provide appropriate and sufficient special information in the model. It is very important that someone who is not involved in the developing of the DWH checks the information for completeness and clarity.

The maintenance of the model is very easy. During the development we had to modify earlier decisions. The relational model and the design of the tool make it very easy to perform the necessary changes and to keep track of the consequences.

All in all building a DWH requires some serious work and we are happy to do it this way.

Contact:

Berthold Weismann
BADENIA Bausparkasse AG
Badeniaplatz 1
D-76114 Karlsruhe

Wilfried Schollenberger
WS Unternehmensberatung und Controlling-Systeme GmbH
Bergstraße 7
D-69120 Heidelberg

Phone: +49 721 995 2915
Fax: +49 721 995 2909

Phone: +49 6221 401409
Fax: +49 6221 401422
EMail: wisch@ibm.net
WEB: www.t-online.de/home/ws-gmbh