Analyzing Educational Process Through a Chain of Data Marts

Viljan Mahnic
University of Ljubljana
Faculty of Computer and Information Science
Trzaska 25, SI-1000 Ljubljana, Slovenia
viljan.mahnic@fri.uni-lj.si

Keywords: data warehouse, data mart, star schema data model.

Received: June 6, 2003

We describe the development strategy, architecture, and logical design of a data warehouse that can be built gradually, exploiting the benefits of the bottom-up, data mart approach. Connections between individual data marts are planned in advance with the aim of building a sequence of data marts that makes it possible to analyze the educational process as a value chain. Queries can be made across different subject areas (viz. enrolment applications, enrolment, examination, and degree records) in order to obtain a snapshot or a slice of the entire value chain that shows how far a subset of students has moved from the enrolment application to their final degree.

1 Introduction

In the early nineties, Bill Inmon introduced the concept of a data warehouse as a subject oriented, integrated, nonvolatile, time variant collection of data in support of management's decisions [5]. To create a data warehouse, data are extracted from different source systems, and then transformed, integrated, and loaded on an appropriate data store. Since then, data warehousing has grown to become one of the most important areas in the information systems field [3]. The benefits of data warehousing are numerous and some organizations are receiving significant returns [12].

The concepts of data warehousing have attracted substantial attention within the EUNIS community. At the EUNIS 1997 Conference, D. Stevenson [9] presented a data warehouse development project from users' and management's perspective, while at EUNIS 1999 M. Bajec et al. [1] proposed to build a data warehouse in order to analyze enrolment applications. At EUNIS 2001, two French initiatives were presented: J-F. Desnos [2] described a comprehensive data warehouse project for French universities, while Flory et al. [4] presented the design and implementation of a data warehouse for research administration. Additionally, the importance of data quality was described in order to ensure successful data warehouse implementation [8].

The aim of our paper is to describe the development strategy, architecture, and logical design of a data warehouse that should provide a unified and integrated source of data for various analyses of educational process at the University of Ljubljana. The main feature of this data warehouse is that it can be built stepwise as a chain of data marts that use common dimension tables, thus providing a suitable architecture for drill-across applications.

2 Development strategy and data warehouse architecture

Even though data warehouses are widespread, there is no common agreement about the best development methodology to use. While Bill Inmon (who is recognized as "the father of data warehousing") recommends a top-down, enterprise data warehouse approach, Ralph Kimball [6, 7] recommends the bottom-up, data mart approach. Using a top-down approach, a global enterprise data warehouse is built first that serves as a basis for implementation of individual data marts. On the other hand, in a bottom-up approach, individual data marts are developed first and later interconnected through common dimensions into a comprehensive data warehouse.

Considering our specific situation (viz. limited budget and the need for tangible results as soon as possible) as well as positive experience reported in the literature [10] we decided to adopt the bottom-up, data mart approach. This approach provides usable data faster, at a lower cost, and with less financial risk. However, in the long term this approach is successful only if all connections between individual data marts are well planned in advance. Therefore, special attention was devoted to logical data warehouse design in order to develop consistent data definitions and define an appropriate structure of the dimension tables that interconnect different data marts.

Additionally, since data warehouses (in general) play an important role in understanding value chains (e.g., by connecting trading partners along the demand or supply chain), we designed our data warehouse with the aim of representing the educational process as a value chain consisting of the following steps:

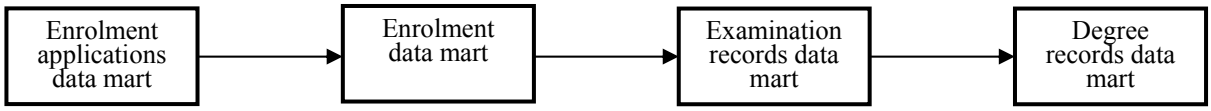

Figure 1: The sequence of the data marts used to analyze the educational process as a value chain

- enrolment application
- first enrolment
- examination
- next enrolment
- examination
- ... (several enrolment and examination steps repeat here)
- degree

In our design, the aforementioned chain is modeled as a sequence of four data marts shown in Figure 1. The first data mart corresponds to enrolment applications, the second one contains enrolment data, the third one corresponds to examination records, and the last one deals with alumni data.

3 Logical design of data marts

Data marts are designed using dimensional modeling introduced by Kimball [6]. Each data mart is represented by a star schema data model (also called star join schema or dimensional model) that is made up of a fact table in the center of the star and several dimension tables as the points of the star. The fact table contains measurable facts that are recorded for each transaction (viz. enrolment application, enrolment, examination, and degree taken, respectively), while dimension tables describe entities (viz. students, study programs, teachers, courses, etc.) that are involved into these transactions.

Each star schema data model can be implemented individually and later integrated with other star schemas through common dimensions as described in Section 4. Figures 2 and 3 represent two sample star schema data models of our data warehouse. Figure 2 describes the logical design of the enrolment data mart that will be implemented first, while Figure 3 represents data in the examination records data mart.

3.1 Enrolment data mart

The fact table
We modelled each enrolment as an event at the intersection of seven dimensions: student, time (viz. academic year), department, study program, year of study, study mode, and type of enrolment (see Figure 2). Such a fact table represents a robust set of many-to-many relationships among these seven dimensions; however, it has only one measurable fact: the fee paid for studies.[1] This means that applications will perform mostly counts. Nevertheless, this table can be queried to answer any number of interesting questions, such as:

- How many students enrolled at each department or study program?
- What is the structure of enrolled students (considering secondary school, profession, secondary school grade, study mode, and/or type of enrolment)?
- Is the number of students (at a particular department or study program) increasing or decreasing?
- What is the progress rate of a particular generation of students?

The student dimension
The student dimension table contains data on students that are enrolled at the university. Each row corresponds to one student and contains his or her personal data. Secondary school, secondary school grade, and profession attributes enable the construction of correlations between students' progress and secondary school, secondary school grades, and secondary school profile of students. On the other hand, zip, county, and region attributes represent a geographic hierarchy that is useful in performing analyses of novice students in connection with the enrolment applications data mart.

The time dimension
Considering the enrolment data mart alone, an explicit time dimension table is not necessary because it is enough to keep the academic year of each enrolment as a degenerate dimension within the fact table. However, since the time dimension is common to all data marts a more elaborate version of time dimension table is necessary that is represented in Figure 3.

The study program dimension
At the University of Ljubljana study programs usually consist of several elective modules which can be further divided into submodules (e.g., in the final year of studies). Therefore, the study program dimension in fact describes individual submodules and defines a useful study program-module-submodule hierarchy that enables the generation of reports with different levels of detail (viz. drill-up and drill-down queries).

The year of study dimension
This is a degenerate dimension since the year of study attribute in the fact table is the only attribute of this dimension. It can be used as a grouping key for pulling together all the students enrolled in the same year of studies.

[1] In Slovenia only part-time students pay fees for their studies.

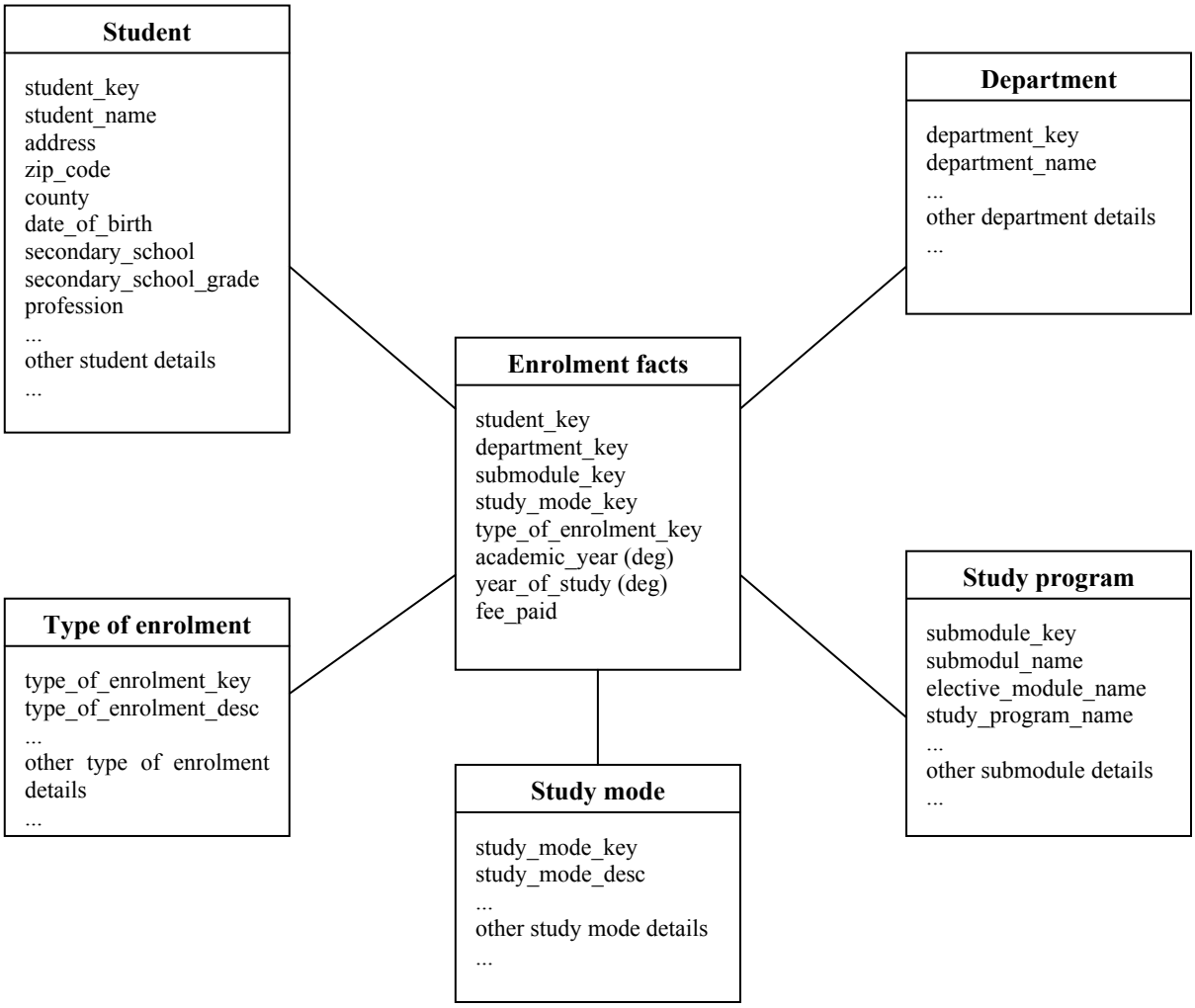

Figure 2: The star schema data model representing the logical design of the enrolment data mart.

The department dimension
The department dimension table describes each member institution. At present, the University of Ljubljana consists of 26 member institutions (22 faculties, 3 academies, and 1 high school).

The study mode dimension
The study mode dimension describes every possible manner of studies (e.g. full-time, part-time, etc.).

The type of enrolment dimension
The type of enrolment dimension describes all possible enrolment types (e.g. first enrolment, repeated enrolment, etc.).

3.2 Examination records data mart

The fact table
Each record of the fact table in Figure 3 corresponds to one examination (viz. the grain of the fact table) and is uniquely defined by a compound key consisting of keys of all dimension tables. There are two measurable facts that can be taken at the intersection of all the dimensions: grade and sequential number of examination attempt.

Dimension tables
Logical design of the examination records data mart comprises six dimension tables through which the data in the fact table can be analyzed: the student dimension, the course dimension, the teacher dimension, the department dimension, the time dimension, and the study program dimension.

The student dimension is the same as in the enrolment data mart. Secondary school, secondary school grade, and profession attributes can again be used in the construction of correlations between examination results and secondary school, secondary school grades, and secondary school profile of students.

Given the fact that the key of the time dimension is simply a date, the time dimension table could be omitted, but we found it useful because additional attributes (such as semester, academic year, and day number overall) enable to slice data by semesters and academic years, as well as to perform simple arithmetic between days across year and month boundaries (e.g., to compute time elapsed from first enrolment till graduation).

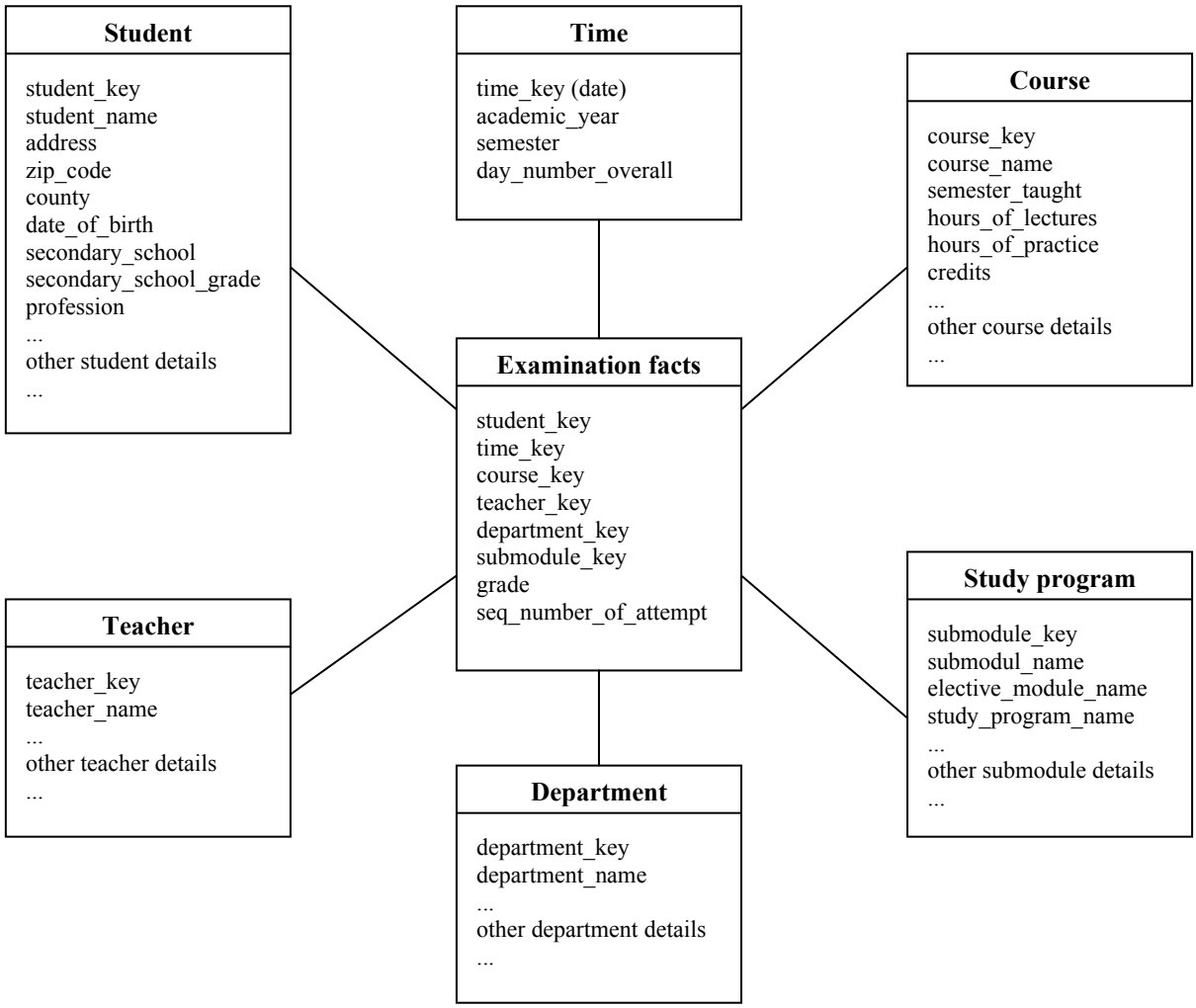

Figure 3: The star schema data model representing the logical design of the examination records data mart.

The course dimension describes every course, and the teacher dimension describes every teacher. The department dimension and the study program dimension are the same as in the enrolment data mart. Using the examination records data mart various analyses of examination results are possible, e.g.:

- What is the average grade of a specified subset of students?[2]
- How many students passed an exam in a given time period at each department or study program?
- What is the average grade and number of examination attempts at given course and/or teacher?
- How do average grades compare among different study programs and/or faculties?

etc.

4 Connecting data marts through common dimensions and drill-across applications

Some dimension tables (e.g., the student dimension, the study program dimension, the time dimension, etc.) are common to all data marts in our data warehouse. These tables can act as "glue" that connects the data marts together and allows meaningful queries to be made across different subject areas (viz. enrolment applications, enrolment records, examination records, and degree records). Using data warehousing terminology, these queries are often called drill-across applications.

In order to support drill-across applications, all constraints on dimension attributes must evaluate to exactly the same set of dimensional entities from one data mart in the value chain to the next data mart in the value chain. For example, a constraint on student dimension at any point in the value chain must mean exactly the same subset of students at all points in the chain. The easiest way to achieve this requirement is to

[2] The subset of students can be specified using attributes in the student dimension table as a source of constraints.

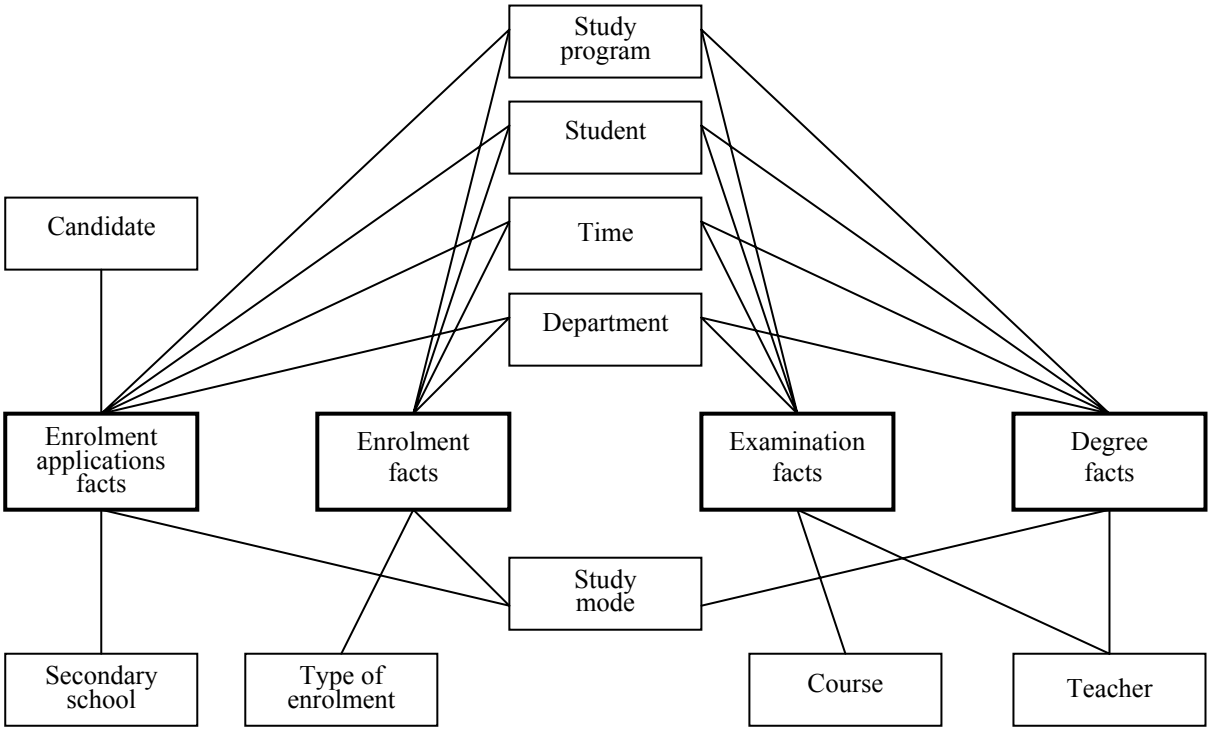

Figure 4: The chain of data marts connected through common dimension tables that are physically implemented only once.

physically implement all common tables only once as shown in Figure 4.[3]

Therefore, it is extremely important to plan the structure of common dimension tables in advance not only to satisfy the needs of individual data marts, but also to provide the necessary connections for drill-across applications. Given the fact that the University of Ljubljana is extremely decentralized and each member institution maintains its own data about students, teachers, and courses, substantial effort was necessary to integrate and cleanse these data in order to build common dimension tables.

Beside dimension tables that have already been described in previous section, there are two additional dimensions shown in Figure 4:

- The candidate dimension table corresponds to all candidates for enrolment. The accepted candidates become part of the student dimension after first enrolment.
- The secondary school dimension describes every secondary school in Slovenia.

[3] There are some special situations when the aforementioned requirement can be achieved without implementing the common dimension only once, e.g. in the case of dimensions with reduced detail and derived dimensions that support aggregates. The interested reader can find more information in [6, pp. 84-85].

5 A sample drill-across report

Using our data warehouse we can imagine that students "move" sequentially through the value chain, and a drill-across report can show a snapshot or a slice of the entire value chain that shows how far a subset of students has moved from the enrolment application to their final degree. For example, using the enrolment applications data mart a subset of candidates that applied for enrolment in a given academic year can be defined. In combination with the enrolment data mart, only those candidates that actually enrolled (viz. became students) can be isolated. For this subset of students, their examination records can be analyzed using the examination records data mart and the number of students who finished their studies can be determined from degree records data mart.

Table 1 represents a sample report obtained by drilling across. Suppose we have several generations of candidates that applied for enrolment at a given department in five consecutive years and we want to track their progress towards the graduation. Given the fact that the department and time dimensions are exactly the same for all data marts, only those data that belong to the department and academic years in question are processed in each data mart. Similarly, the common student dimension assures that the same subset of students is taken into account in the whole value chain. Using the enrolment applications data mart the exact number of applicants and the number of approved

Department XXX

Academic Year	Applicants	Approved	Enrolled	Average Grade	Exams Passed	Graduated	Length of Studies
1992/93	237	162	150	8.16	29.3	77	6.15
1993/94	199	165	143	8.02	27.1	71	6.32
1994/95	182	158	144	7.57	24.3	64	6.87
1995/96	176	151	145	7.03	19.5	56	7.02
1996/97	183	154	150	7.43	22.4	57	6.21

Table 1: A sample drill-across report

Department XXX

Academic Year	Study Program	Applicants	Approved	Enrolled	Average Grade	Exams Passed	Graduated	Length of Studies
1992/93	AAA	125	101	95	8.21	30.1	52	6.08
	BBB	112	61	55	8.07	27.9	25	6.32
Total 1992/93		237	162	150	8.16	29.3	77	6.15
1993/94	AAA	121	110	96	8.12	28.1	52	6.12
	BBB	78	61	47	7.81	25.0	19	6.89
Total 1993/94		199	165	143	8.02	27.1	71	6.32
1994/95	AAA	119	108	94	7.78	27.1	47	6.65
	BBB	63	50	50	7.18	19.0	17	7.47
Total 1994/95		182	158	144	7.57	24.3	64	6.87

etc.

Table 2: Report refinement by drilling down

applications can be determined. The enrolment data mart enables the computation of the actual number of enrolled students, while the examination records data mart provides the average grade[4] and average number of exams passed for these students. Finally, the degree records data mart is used to compute the number of graduates in each generation as well as the average length of their studies (in years).

By drilling down we can still refine our query in order to obtain more detailed information about each generation of students. By simply adding the study program name as a new row header the same analysis can be obtained for each study program separately (see Table 2). We can further subdivide the subset of students in each generation by using other attributes from the study program hierarchy (viz. elective module name and submodule name) as well as by choosing row headers from dimensional attributes of other dimensions (e.g., secondary school, secondary school grade, secondary school profession).

6 Conclusions

We described the design of a data warehouse that can be implemented gradually as a chain of data marts connected through common dimension tables. Each data mart was represented using the star schema data model and a special attention was devoted to the definition of common dimensions in order to enable drill-across applications. The proposed design describes the educational process as a value chain and makes it possible to analyze how far a subset of students has moved from the enrolment application to their final degree.

[4] In Slovenia the following grades are used: 1 to 5 – insufficient; 6 – sufficient; 7 – good; 8, 9 – very good; 10 – excellent.

References

[1] Bajec, M., Rupnik, R., Krisper, M. Using Data Warehouses in University Information Systems, in K. Sarlin (ed.) EUNIS 99 – Information Technology Shaping European Universities, pp. 115-121.

[2] Desnos, J-F. A Data Warehouse for French Universities, Informatica, Vol. 25, No. 2, July 2001, pp. 177-181.

[3] Eckerson, W.W. Evolution of Data Warehousing: The Trend toward Analytical Applications, Boston, MA: The Patricia Seybold Group (April 28, 1999), pp. 1-8.

[4] Flory, A. et al. Design and Implementation of a Data Warehouse for Research Administration Universities, in J. Knop and P. Schirmbacher (eds.) EUNIS 2001 – The Changing Universities, The Role of Technology, Berlin, March 2001, pp. 164-167.

[5] Inmon, B. Building the Data Warehouse, QED Publishing Group, 1992.

[6] Kimball, R. The Data Warehouse Toolkit, John Wiley & Sons, 1996.

[7] Kimball, R. et al. The Data Warehouse Lifecycle Toolkit, John Wiley & Sons, 1998.

[8] Mahnic, V. Rozanc, I. Data Quality: A Prerequisite for Successful Data Warehouse Implementation, Informatica, Vol. 25, No. 2, July 2001, pp. 183-188.

[9] Stevenson, D. Data Warehouse and Executive Information Systems – Ignoring the Hype, in J-F. Desnos and Y. Epelboin (eds.) European Cooperation in Higher Education Information Systems, Grenoble, France, September 1997, pp. 202-207.

[10] Watson, J.H. et al. Sherwin-Williams' Data Mart Strategy: Creating Intelligence across the Supply Chain, Communications of the Association for Information Systems, Vol. 5, Article 9, May 2001.

[11] Watson, J.H. Recent Developments in Data Warehousing, Communications of the Association for Information Systems, Vol. 8, 2001, pp. 1-25.

[12] Watson, J.H. et al. The Benefits of Data Warehousing: Why Some Organizations Realize Exceptional Payoffs, Information & Management, Vol. 39, 2002, pp. 491-502.