

Data Quality: A Prerequisite for Successful Data Warehouse Implementation

Viljan Mahnic and Igor Rozanc

University of Ljubljana Faculty of Computer and Information Science

Viljan.Mahnic@fri.uni-lj.si

Igor.Rozanc@fri.uni-lj.si

www.fri.uni-lj.si

Trzaska 25, SI-1000 Ljubljana, Slovenia

Keywords: data warehouse, information quality, total quality data management, data quality assessment

Abstract: Building a data warehouse for a large decentralized university such as the University of Ljubljana is an attractive challenge, but also a risky and demanding task. Experience has shown that projects attempting to integrate data are especially vulnerable to data quality issues. Therefore, before embarking on a data warehouse initiative a thorough quality assessment of the source data is necessary. We describe how the assessment criteria based on the Total Quality data Management Methodology were adapted to our specific needs and used to determine the quality of student records data at two member institutions, viz. the Faculty of Computer and Information Science, and the Faculty of Electrical Engineering. The most important results of the assessment are described and proposals are given for further activities. The assessment has shown that the student records data at the Faculty of Computer and Information Science and Faculty of Electrical Engineering are good enough to be used as source for the global warehouse at the university level after some data cleansing takes place. Additionally, special attention must be devoted to the integration of such data that are replicated at many individual departments (viz. employees, subjects taught, and students). Therefore, we propose that a unique coding scheme for all employees, students, and subjects taught be defined in the first step of the data warehouse design, and an ongoing data quality management process is established clearly defining the roles and responsibilities of all personnel involved.

Introduction

The University of Ljubljana is the largest university in Slovenia. It consists of 26 member institutions (20 faculties, 3 academies, and 3 colleges) and has more than 40,000 students. In the past, the member institutions had substantial autonomy regarding the usage of information technologies, which led to uncoordinated development of their information systems. Different applications were developed for the same purpose and due to this heterogeneity it is very difficult or even impossible to create reports that require cross-referencing data from different institutions or application areas.

In such a situation, the building of a data warehouse at the university level seems to be the most appropriate solution. Therefore, a pilot project started with the aim of defining a data model for a global data base which will be fed by data from member institutions on a regular basis and which will serve as a basis for analytical processing at the university level. One of the main tasks within this project is to define the granularity of data in the data warehouse and different levels of detail which will support best the decision making processes.

However, experience from other organizations has shown that projects attempting to integrate data are especially vulnerable to data quality issues [1]. A recent study by the Standish Group states that 83 percent of data migration projects overrun their budget (or fail) primarily as a result of misunderstandings about the source data and data definitions. Similar surveys conducted by the Gartner Group point to data quality as a leading reason for overruns and failed projects [2].

In order to avoid such a pitfall, a thorough assessment of the quality of data that are used as input to the global data warehouse is necessary. Since we decided that our data warehouse will be populated gradually, starting with student records data, the quality of these data was analyzed first. The aim of this paper is to describe in detail the assessment methodology and results obtained at two typical member institutions, viz. the Faculty of Computer and Information Science (FCIS), and the Faculty of Electrical Engineering (FEE).

The assessment criteria were defined according to English's Total Quality data Management methodology [3] and Forino's recommendations [2]. A description of these criteria is given in Section 2, while the results of the assessment are presented in section 3. Section 4 describes the proposals for further activities, and section 5 summarizes the most important conclusions.

Data Quality Assessment Methodology

Data quality assurance is a complex problem that requires a systematic approach. English [3] proposes a comprehensive Total Quality data Management Methodology (TQdM), which consists of 5 processes of measuring and improving information quality, and an umbrella process for bringing about cultural and environmental changes to sustain information quality improvement as a management tool and a habit:

Process 1: Assess Data Definition & Information Architecture Quality

Process 2: Assess Information Quality

Process 3: Measure Nonquality Information Costs

Process 4: Reengineer and Cleanse Data

Process 5: Improve Information Process Quality

Process 6: Establish the Information Quality Environment

Each process is further divided into steps that must be followed in order to achieve the desired data quality. Organizations embarking on data warehouse initiatives and that do not yet have an information quality function must conduct many of these steps, but may do so in a different sequence, based on their specific needs.

Considering our specific needs, we concentrated on Process 2 which defines two aspects of information quality: the inherent information quality and the pragmatic information quality. Inherent infor-

Faculty	No. of files	No. of checked attributes	No. of checked records	No. of records with missing values	% of erroneous records
FCIS	33	69	201617	43	0.021 %
FEE	33	69	405436	996	0.246 %
TOTAL	66	138	607053	1039	0.171 %

Table 1: Completeness of vaues (Summary Data)

Field name	Field description	File name	Total no. of records	No. of records with missing values	% of erroneous records
VPIS_ST	student_id	DIPLOMA	961	18	1.873 %
IME	employee_ first - name	DELAVEC	391	4	1.023 %

Table 2: Fields with most missing values (Faculty of Computer and Information Science)

Field name	Field description	File name	Total no. of records	No. of records with missing values	% of erroneous records
VPIS_ST	student id	DIPLOMA	2926	185	6.323 %
IME_D	employee_ first - name	DELAVEC	399	4	1.003 %
DELAVEC	employee_id	VAJE	152071	737	0.485 %

Table 3: Fields with most missing values (Faculty of Electrical Engineering)

mation quality is the correctness or accuracy of data, while pragmatic information quality is the value that accurate data has in supporting the work of the enterprise. In this paper the results of inherent information quality assessment are described.

In order to determine the quality of data a field-by-field assessment is required. However, simply having data is not enough, but the context for which the data is to exist must also be known. To put in other terms, a clear data definition or the so called meta data must be provided [2]. Generally speaking, one can find meta data in data models, data dictionaries, repositories, specifications, etc. If current meta data does not exist, then a subject matter expert is needed, and meta data is a much-desired by-product of a data quality assessment.

The extent of assessment in great deal depends on the availability of meta data. According to [2], assessments usually focus on one or more of the following types of quality criteria:

1. Data type integrity
2. Business rule integrity
3. Name and address integrity

If the assessment team knows nothing more than field names, types and sizes, then the focus is on testing the field's integrity based on it's type (numeric, alphanumeric, date, etc.). If additional characteristics of the field are provided (domain, relationship with other fields, etc.), then business rule integrity is also performed. Finally, if name and address data is critical (particularly if it will be consolidated with other data), then name and testing should be performed.

On the other hand, English [3] defines the following inherent information quality characteristics and measures:

1. Definition Conformance
2. Completeness (of values)
3. Validity, or business rule conformance
4. Accuracy to surrogate source

5. Accuracy (to reality)
6. Precision
7. Nonduplication (of occurrences)
8. Equivalence of redundant or distributed data
9. Concurrency of redundant or distributed data
10. Accessibility

Considering the aforementioned quality characteristics and our specific needs we decided to measure the quality of our data using the following criteria:

1. **Completeness of values:** Users were encouraged to prioritize a subset of source fields that must have a nonnull value. For each field the percentage of records with missing values was computed.
2. **Validity, or business rule conformance:** For a chosen subset of most important fields we measured the degree of conformance of data values to their domains and business rules.
 - a. All 1:N relationships were examined for existence of foreign keys in master tables.
 - b. Since time represents an important dimension in a dimensional data warehouse a range check was performed on all date fields in order to find possible out-of-range values.
 - c. A range check was performed on all other fields from the highest priority subset (e.g. academic year, year of study, grades etc.).
 - d. Special cross-checks were defined for fields having relationships with other fields that further define the allowable domain set, e.g.

Faculty	No. of files	No. of checked relationships	No. Of checked records	No. of Relationships with Errors	No. of Records with Errors	% of Relationships with Errors	% of Records with Errors
FCIS	49	207	1361758	38	6360	18.357 %	0.467 %
FE	49	207	2702005	51	16605	24.638 %	0.615 %
TOTAL	98	414	4063763	89	22965	21.498 %	0.565 %

Table 4: Existence of foreign keys in master tables (Summary Data)

Entity (File)	No. of records	Foreign key field	No. of non-existent values	% of errors	Name
Description		DIPLOMA	961	NACIN	type of study code
546	56.816 %	STUD_F	4062	OBCINA_SS	territorial unit code of student's secondary school
1340	32.989 %	STUD_F	4062	OBCINA_R	territorial unit code of student's place of birth
1331	32.767 %	STUD_F	4062	OBCINA_S	territorial unit code of student's permanent residence
1314	32.349 %	STUD_F	4062	OBCINA_Z	territorial unit code of student's temporary residence
603	14.845 %				

Table 5: Problematic relationships (Faculty of Computer and Information Science)

Entity (File)	No. of records	Foreign key field	No. of non-existent values	% of errors	Name
Description		DIPLOMA	2926	NACIN	type of study code
1461	49.932 %	STUD_F	8164	OBCINA_SS	territorial unit code of student's secondary school
3565	42.908 %	STUD_F	8164	OBCINA_R	territorial unit code of student's place of birth
3503	42.896 %	STUD_F	8164	OBCINA_S	territorial unit code of student's permanent residence
3502	42.896 %	STUD_F	8164	OBCINA_Z	territorial unit code of student's temporary residence
1422	17.418 %				

Table 6: Problematic relationships (Faculty of Electrical Engineering)

File	No. of records	No. of duplicated primary keys	% of errors
VSI.DBF	1314	68	5.175 %
SS_POKLI.DBF	1566	42	2.682 %
CENTRI.DBF	51	1	1.960 %
ZAVOD.DBF	100	1	1.000 %

Table 7: Nonduplication of primary keys (Code tables maintained by the University computing center)

File	No. of records	No. of duplicated primary keys	% of errors
IZJEME.DBF	56	1	1.786 %
TEMA.DBF	978	3	0.307 %
DELAVEC.DBF	391	1	0.226 %
DVIG.DBF	922	1	0.108 %

Table 8: Nonduplication of primary keys (FCIS)

File	No. of records	No. of duplicated primary keys	% of errors
SPP.DBF	270	11	4.074 %
PRED_PR.DBF	746	8	1.072 %
PP.DBF	2557	12	0.469 %
DVIG.DBF	2825	9	0.319 %

Table 9: Nonduplication of primary keys (FEE)

- when a student applies for an exam for the first time (i.e. NO_OF_ATTEMPTS=1) the date of previous examination must be blank and vice versa
- the date of previous examination must be lesser than the date of next examination
- in each degree record the difference between the degree date and thesis issue date must be positive and less than six months.

Nonduplication of occurrences and equivalence of redundant and distributed data: In our case two different tests were performed:

- All files in the student records database were checked for eventual duplications of primary keys.
- The student and employee files at both faculties were checked for the existence of the records that are duplicate representations of the same student or employee respectively.

Assessment Results

Completeness of values

The most important fields in 33 files were checked for nonnull values, and it was found, that only few values were missing. Results are summarized in Table 1, while Tables 2 and 3 show the fields and files with most missing values for each faculty, respectively. The greatest problem represent the missing student_id values in alumni records. This is a consequence of the fact that 15 years ago (when alumni records started) some students were not assigned a student identification number.

Validity, or business rule conformance

Existence of foreign keys in master tables: The examination of all 1:N relationships revealed that in about a half percent of records foreign keys do not have their counterpart values in the corresponding master tables (see Table 4). However, this average is misleading, since the majority of errors appear within a small number of relationships. The following relationships appeared to be the most problematic at both faculties (see Tables 5 and 6):

1. the relationship between entities DIPLOMA (student's degree thesis) and NACIN (a code table of possible types of study, e.g. full-time or part-time);
2. multiple relationships between entities STUD_F (student's personal data) and OBCINA (a code table of territorial units in Slove-

Entity (File)	Description	No. of replications	same primary key	different primary key
DELAVEC	Employees	388	0	STUDENT
Students	4	121		

Table 10: Equivalence of redundant occurrences (FCIS and FEE)

nia) representing the territorial unit of student's residence, place of birth, secondary school finished etc.

The first of the aforementioned problems is simply a consequence of the fact that (within the alumni records database) the faculties started collecting the type of study data in 1994, while this datum is missing in older records. The second problem is much harder and was caused by significant changes in territorial organization at the local level in Slovenia after independence, which required several consequent modifications of the corresponding code table. Although each year's enrolment data corresponded to the currently valid version of the code table, there are a lot of inconsistencies when we look at the data in time perspective.

Range checking of date fields did not reveal any serious problems. In the worst case the rate of out-of-range field values reached 0.149 % at FCIS, and 0.493% at FEE.

Range checking of other fields from the highest priority subset also yielded quite good results. The field IME_D (viz. employee's first name) was ranked the worst at both faculties containing 1.023 % erroneous values at FCIS, and 1.003 % erroneous values at FEE.

Special cross-checks pointed out that the business rule requiring that each student completes his/her degree project in six months is sometimes violated. Namely, the difference between the degree date (datum zagovora) and thesis issue date (datum izstavitve teme) was more than six months in 3.362 % of cases at FCIS, and in 4.248 % of cases at FEE. Error rates reported by other cross-checking tests were less than 1 %.

Nonduplication of occurrences and equivalence of redundant and distributed data

Nonduplication of primary keys: The student records information system at FCIS and FEE [4] (as well as at other member institutions of the University of Ljubljana) is implemented using Clipper which does not automatically force the uniqueness of primary keys. In spite of the fact that all programs have been carefully written in order to avoid duplicates, some can be introduced through the manual maintenance of data, especially code tables. Therefore, all files in the student records database were checked for eventual duplications of primary keys. An excerpt of assessment results showing only files with most duplicates is presented in tables 7 through 9. Table 7 shows the number of duplicated primary keys in code tables that are maintained centrally by the University computing center for all member institutions, while tables 8 and 9 refer to duplicates at FCIS and FEE respectively. A relatively high percentage of errors in some code tables indicates that the maintenance of code tables at the University computing center should be improved.

Equivalence of redundant and distributed data: Given the fact that the same teachers teach at both faculties as well as that some students study at both faculties (e.g. a student can obtain his/her B. Sc. degree at FEE and enrolls for graduate study at FCIS or vice versa) some data are replicated across faculties. In order to state the extent of such a replication two measures were introduced: the number of replicated entity occurrences (viz. the same teacher or student in both databases) with the same primary key, and the number of replicated entity with different primary keys.

Assessment revealed that employee files at both faculties are consistent: all replicated employees have the same primary key. This is not the case with student files. Due to the decentralized organization of the University of Ljubljana each faculty assigns its students a different identification number regardless the fact that the student has already been enrolled at another faculty. Unfortunately, these kind of inconsistencies may be a source of major problems when integrating data into a global warehouse.

Proposed Further Actions

On the basis of assessment results we propose two kinds of further actions: cleansing of source data and an appropriate design of the global data warehouse at the university level.

Some erroneous source data can be cleansed automatically (e.g. missing type of study code in file DIPLOMA), while other data require manual or combined manual and automatic approach (e.g. removal of duplicated primary keys, re-establishment of relationships using territorial unit codes). Some errors (e.g. out-of-range values) can be prevented by the incorporation of appropriate controls in the program code.

Special attention must be devoted to the maintenance of code tables that are common for the whole university. A relatively high percentage of duplicate codes in code tables, maintained by the university computing center up to now indicates that the maintenance of these code tables must improve.

Additional problems could arise during the integration of those data and code tables that are at present maintained by individual departments (viz. employees, subjects taught, and students). Although our assessment did not reveal serious problems within each department, many duplications and code conflicts may occur during the integration. Therefore, we propose that a unique coding scheme for all employees, students, and subjects taught is defined in the first step of the data warehouse design.

The data warehouse design must be based on principles of TQDM methodology. Data standards must be defined and data definition and information architecture quality assessment must take place before programming and population of the data warehouse begins. An ongoing data quality management process must be established and the roles and responsibilities of a data quality administrator, subject area champions, data oversight committee, and data owners clearly defined [4].

Conclusions

Our paper was intended to increase the awareness of the importance of data quality not only when building a data warehouse but also in operational environments that support transactional processing. An assessment methodology to empirically determine the data quality was described and the results of the assessment were presented.

Considering the assessment results we estimate that source data at the Faculty of Computer and Information Science and Faculty of Electrical Engineering are good enough to be used as source for the glo-

bal warehouse at the university level after some data cleansing takes place. In first place, missing student identification numbers must be provided in alumni records and the broken relationships using territorial units codes must be re-established in students' personal data. and the maintenance of common code tables at the University computing center must improve.

During the design of the global data warehouse a special attention must be devoted to the integration of those data that may be replicated at many individual departments (viz. employees, subjects taught, and students). Since each department has its own policy of coding, many duplications and code conflicts may occur during the integration. Therefore, we propose that a unique coding scheme for all employees, students, and subjects taught is defined in the first step of the data warehouse design. Additionally, an ongoing data quality management process must be established and the roles and responsibilities of all personnel involved should be clearly defined.

Bibliography

- [1] Celko, J., McDonald, J., Don't Warehouse Dirty Data, Datamation, October 15, 1995, pp. 42 - 53.
- [2] Farino, R., The Data Quality Assessment, Part 1, DM Review Online, August 2000, <http://www.dmreview.com>
- [3] English, L.P., Improving Data Warehouse and Business Information Quality, John Wiley & Sons, Inc., 1999, ISBN 0-471-25383-9.
- [4] Mahnic, V., Vilfan, B., Design of the Student Records Information System at the University of Ljubljana, Proceedings of the EUNIS'95 Congress, Düsseldorf, Germany, November 1995, pp. 207 - 220.
- [5] Kachur, R., Data Quality Assessment for Data Warehouse Design, DM Review Online, April 2000, <http://www.dmreview.com>