Data Policy for Increasing the Data Quality in Intelligent Systems

Tatjana Welzer[1], Izidor Golob[1], Marjan Družovec[2], Boštjan Brumen[1]

[1]University of Maribor
Faculty of Electrical Engineering and Computer Science
Smetanova 17, Si-2000 Maribor, Slovenia
[2]University of Maribor
Faculty of Mechanical Engineering,
Smetanova 17, Si-2000 Maribor, Slovenia

{ welzer | izidor.golob | marjan.druzovec | bostjan.brumen }@uni-mb.si,

Abstract

Use of data in various areas and their electronic availability has put importance of data quality to higher level. In general data quality has syntactic and semantic component. The syntactic component is relatively easy to achieve if supported by tools, while semantic component requires more research. In many cases such data come from different sources, which are distributed across enterprise and are at different quality levels. Special attention needs to be paid to data upon which critical decisions are met in/for intelligent systems. In the present paper we will focus on the semantic component of data quality in the selected domain and data policy for increasing the quality of data used and/or acquired in intelligent systems.

Keywords: *information and data quality, data policy, intelligent systems, reuse*

1 Introduction

Many of approaches to improve information quality have been developed over the years and employed in various situations. Most of these approaches are only vaguely aware that for the information quality data quality is prerequisite.

Data are of high quality if they are fit for their uses in information/intelligent systems. Data are fit for use if they are free of defects and possess desired features (Redman, T.C. 2001). In the information technology aggressive steps to improve the data quality are being taken.

In our contribution, we will concentrate on the problem of data quality in the medical environment. The medical environment was selected as an experimental domain for presenting problems of data quality and benefits of data policy including its influence on higher quality of data. The main research is concentrated on medical diagnostics.

A demand for a diagnostics is presented in a wide variety of human activities; from diagnostics of technical systems, medical diagnostics, to finding incorrect hypotheses in scientific theories. The diagnostic problem is to determine those components of the system for which we assumed to be functioning abnormally, what means the discrepancy between the observed and the correct system behavior (Welzer, T., Brumen, B., Golob, I. and Družovec, M. 2002).

The data quality (syntactic as well as semantic component) has extremely huge impact on the diagnostic process, especially in the medical environment. Independent how the diagnostic procedure is done, data input has a direct influence on the diagnoses, their explanations, questions posed to the user and the solutions. Either semantic or syntactic incorrectness can have fatal consequences by making decision, not only in medical environment.

2 Quality of data

The quality of data limits the ability of the end user to make correct decisions, which can have fatal consequences. There are a number of indicators of quality of data: accuracy, integrity, consistency, accessibility, comprehensives, timeliness and completeness, among others. The data must follow business rules and be free of anomalies. While being a subjective measure, the user's satisfaction with the quality of the data and the information derived from it, is arguably the most important indicator of them

all (Tayi, G.K. and Ballou, W 1998 and Welzer, T., Brumen, B., Golob, I. and Družovec, M. 2002).

There are many reasons why it is difficult to capture and maintain quality data. Some of the difficulties are process-related, some are human-related problems, and yet other difficulties are technology-related problems. All of the problems have a consequence in bad data, being it semantically or syntactically bad. Usually these two components are inter-related and inconsistence.

Process-related problems are frequently caused by the user entering the data into an operational system at the wrong point of the business process or by lack of understanding of the meaning of the data (Welzer T., Brumen B., Golob I., and Sanchez J.L., Družovec M. 2004). Difficulties with employees entering incorrect data into systems can be decreased by changing the emphasis on pure speed of processing to quality of processing, where quality is composed of both speed and accuracy. A technical reason that is causing several problems is the lack of architecture for the system, which can, for example, lead to systems where the rules about the data are embedded in the legacy code rather than being part of the data. As a concrete example, a computer program might use the same field to store data of different meaning, depending on the value of data. While conserving the amount of storage space required, it is very difficult to verify the correctness of the data entered.

Regardless of the cause of the problems, it is important to identify the source of the problem, analyze its impacts and, where possible, propose a solution. By this we have to be aware of the fact that there are usually many customers with different needs and some needs even conflict. Additionally the customer needs change all the time and what was good enough one day (fit the data quality) is simply not good enough at the next day (do not feet the data quality) (Redman, T.C. 2001).

Data quality issue is especially important in data warehouses, data mining, intelligent systems and all other sensitive areas. The introduction of a data warehouse and data mining (intelligent systems) increased the priority of data quality as the risks and costs of inadequate quality become more visible and more real, and after all, more costly (English, L.P. 1999). One of the most important features of a data warehouse is the ability to integrate data from many legacy systems including intelligent systems. The problem of poor data quality is one of the most difficult problems to solve while constructing a data warehouse. Poor data quality can stymie successful implementation of data warehouse. The main goal of course is to build data quality into the warehouse, rather than worrying about it after implementation (Redman, T.C. 2001).

Because of bad data quality, more time and money is spent than anyone estimates initially. Pyle (Pyle, D. 1999) makes an approximate estimate that the data preparation sub-process can take up to 90% percent of the time and money available for the whole knowledge discovery process and/or data mining.

Unless organizations investigate each and every instance at the data value level and then reengineer and consolidate the data prior to migration to a data warehouse, they will naively flood the new structures with erroneous, inaccessible, and improperly integrated data. As a result, the decision makers and data miners will be unable to attain an accurate, consolidated picture of their business. If the consolidated data are wrong, the decisions may be wrong as well.

3 Data quality specifics in taking decisions for specific environment

Medical environments require special information systems because of their dual nature. First, the business part is required, which is just as complex as a business part in any other enterprise. Second, the medical part needs to be interwoven with the business part. In a general business system, we usually do not have such a duality. The medical part has its own specifics and requirements and if it is generally possible to (easily) reuse business objects (data models, applications) from different enterprises for the business part of a medical system (Rine, D.C. 1997), we cannot say the same for the medical part. As mentioned, it is essential that the medical part is tightly connected to the business part.

Decisions in a medical environment are made at two levels. The first level is a management level, which decides upon regular business (such as hiring workforce, finding best suppliers, taking care of stock, and others). The second level is a medical level, where the trained professionals decide upon treatments of patients. Clearly, the latter decisions depend on former, and vice versa.

Decisions made in a medical environment are very sensitive because they affect the “business object” (a patient) directly. Any decision, being it managerial or medical, can have fatal and devastating consequences. For this reason, the real world (medical diagnostic in our case) has to be – first of all – correctly modeled. Second, the data need to be correct, adequate, and available. Moreover, the data should provide a firm foundation

for information retrieval for decision-making process and further for knowledge discovery.

From data quality point of view, we have to assure quality in each of the sub-systems. Additionally, the final (integrated) medical system, composed of business and medical part, needs to be data quality validated again. The integrated high quality systems (intelligent systems) are the desired goal of system designers. To reach the goal the following steps have to be followed:

- Separate data issues from more traditional technical issues and assign lead responsibility for data to someone within the community (separate responsibility for management level and medical level in the case of medical environment is requested).
- As with all quality efforts the needs have to be understood correctly.
- Already existing data have to be check up again. The importance of additional checking is growing with the sensitiveness of domains (medical environment).

In the case of the medical systems, integration may be easier in some parts due to unique identifier each patient has. Usually, this is a social or health security number. Without a unique identifier to match the records, some other less reliable methods must be used to match the records.

We want to stress that an organization sometimes may be better off not having certain data (responsibility of community, request for additional check) than having inaccurate date, especially if those relying on the data are not aware of its inaccuracy. For example, a hospital would be better off not knowing a patient's blood type than wrongly believing it to be O+. A problem of semantic data quality is evident. How is such a problem to be solved?

One of the possible solutions is to introduce data policy.

4 Data policy

To satisfy the quality requests in medical environments for needs of information/intelligent systems, data warehouse and/or data mining, specific requirements have to be fulfilled (Chapter 3). Those requirements based on domain specific, present actually quite general requests that can be written down as data policy. A policy is a plan, course of actions or set of rules intended to influence and determine decisions, actions and other matters (Whitham, M.E. and Mattord 2003)

The origin of defining data policy is the fact that the responsibility for the quality of data has to be assigned to those who create the data or to those who are as close to data creation as possible. That means that the data policy supports the work of those people.

With the goal to define easily used but powerful policy we are suggesting the following structure:

- Introduction
  - Purpose
  - Audience
  - Definitions
  - Related work
  - Basic approach
  - Responsibility
- Data policy
  - What is data policy and why have one
  - What makes a good data policy
- Structure
  - Objectives
  - Categories of data policy
    - Domain rules
    - Syntactic data quality
    - Dimensions of data quality
      - Relevance
      - Clarity of definition
      - Comprehensivenes
      - …..
- Policy management
  - Responsible for reviews
  - Schedule of reviews,
  - Recommendation for reviews
  - Policy issuance and revision date.

According to the suggested structure, are of our further interest only dimensions of data quality (availability, security, comprehensiveness, flexibility, appropriate use, obtainability, semantic consistency, simplicity, relevancy, completeness, consistency, portability, naming, relevancy, occurency, definitions, robustness, homogeneity, redundancy,...). Further we set out especially those dimensions (see the policy structure), which assure data quality by considering conceptual models and/or parts of them:

- Relevance - objects needed by the applications are included in conceptual models.
- Clarity of definition - all the terms used in the conceptual model are clearly defined.

- Comprehensiveness - each needed attribute should be included.
- Occurrence identifiably - identification of the individual objects is made easy.
- Homogeneity, structural consistency - object level enables the uniformity of stored concepts.
- Minimum redundancy - only checked conceptual models are included.
- Semantic consistency - conceptual models are clear and organized according to the application domains.
- Robustness, flexibility - through the reuse both characteristics are fulfilled.

5 Conclusion

The medical diagnostic is a form of knowledge discovery from data. It is a kind of intelligent system. Not only that the data need to be accurate, valid and timely (the semantic component of the data quality), but also the structures from which we obtain the needed data to be valid and syntactically correct.

Too often the structures are neglected or taken for granted. Namely, if the information or knowledge obtained from the data does not satisfy users' wishes and needs, the diagnostic process needs to be reverted to the preceding steps – obtaining the data. But this does not lead to the final destination, since the structures are not suitable for the task. We argue that the data (information, knowledge) quality heavily depends on the data policy which structure is introduced in the paper.

In the paper we presented the diagnostic process from the data quality point of view, exposing both, the semantic and the syntactic component. They can both be improved at the same time if the structures that hold the data are checked for the data policy criteria - the syntactic data quality and dimensions of data quality.

6 References

Welzer, T. and Rozman, I. (1998): Information Quality by MetaModel. *In Proceedings of Software Quality Management VI. Quality improvement issues.* 81-88. HAWKINS. C (eds). Springer. London.

Welzer, T., Brumen, B., Golob, I. and Družovec, M. (2002): Medical diagnostic and data quality. *In Proceedings of 15th IEEE symposium on computer-based medical systems*. 97-101. KOKOL P., STIGLIC B., ZORMAN M. and ZAZULA, D. (eds) IEEE Computer society. Los Alamitos.

Welzer, T. and Družovec, M. (2000): Similarity search in Database Reusability – a Support for efficient design of conceptual models. *In Contemporary Applications and Research Issues in Industrial Product Moddeling*. 23-34. HELLO. P. and WELZER. T. (eds). University of Vaasa.

Freeman, P. (1987): Reusable Software Engineering Concepts and Research Directions. *In IEEE Tutorial Software Reusability.* FREEMAN. P. (ed.). IEEE.

English, L.P. (1999): *Improving data warehouse and Business Information Quality*. John Wiley & Sons.

Pyle, D. (1999): *Data Preparation for Data Mining*. Morgan Kaufmann Publishers, Inc., San Francisco, California, USA.

Redman, T.C. (1996): *Data Quality for the Information Age*. Artech House.

Redman, T.C. (2001): *Data Quality, The field Guide*. Digital Press, Boston, USA.

Whitham, M.E. and Mattord (2003): *Principles of information Security.* Thomson. Canada.

Reiter, R. (1987): A Theory of Diagnosis from First Principles. *Artificial Intelligence* **3**(2):57-95.

Rine, D.C. (1997): Supporting Reuse with Object Technology. *IEEE Computer*, **30**(10): 43-45.

Tayi, G.K. and Ballou, W (1998): Examining Data Quality. *Communications of the ACM*, **4**(12):54-57.

Welzer T., Brumen B., Golob I., and Sanchez J.L., Družovec M. (2004): *Diagnostic process from the data quality point of view*, Journal of Medical Systems, Kluwer (will be published).