TECHNOLOGIES FOR ENSURING DATA QUALITY AND SECURITY IN INDUSTRIAL INFORMATION SYSTEMS

Yordanka Anastasova

University of Mining and Geology “St. Ivan Rilski”, 1700 Sofia; yordanka.anastasova@mgu.bg

ABSTRACT. There are different definitions regarding the quality of data in information systems, however, no absolute quality criteria could be set applicable to the different types of systems. From the point of view of information technology, data quality is defined as a set of qualitative or quantitative criteria. Key features of data quality are accuracy, completeness, consistency, uniqueness, relevance and timeliness. Using data in information systems based on client-server technology also emphasises accessibility and especially data security. In industrial information systems, the data is considered to be of high quality if it sufficiently reflects the described object and can be used to make effective management decisions. This article explores different technologies to provide the optimal set of required features that guarantee the quality of data in information systems applicable in the industry.

Keywords: data quality, security, technology, industrial information systems

ТЕХНОЛОГИИ ЗА ОСИГУРЯВАНЕ КАЧЕСТВОТО И СИГУРНОСТТА НА ДАННИТЕ ПРИ ИНДУСТРИАЛНИ ИНФОРМАЦИОННИ СИСТЕМИ

Йорданка Анастасова

Минно-геоложки университет ”Св. Иван Рилски”, 1700 София

РЕЗЮМЕ. Съществуват различни дефиниции относно качество на данните в информационните системи, като не може да бъдат определени абсолютни критерии за качеството им, валидни за различните видове системи.

От гледна точка на информационните технологии качеството на данните се дефинира като набор от качествени или количествени критерии. Основни характеристики на качеството на данните са точност, пълнота, последователност, уникалност, приложимост и своевременност. При използване на данните в информационните системи, базирани на технологията клиент-сървър се акцентира и на достъпността и особено на сигурността на данните.

В индустриалните информационни системи данните се считат за висококачествени, ако достатъчно реално отразяват описвания обект и служат за вземане на ефективни управленски решения.

Настоящата статия разглежда различни технологии за осигуряване на оптималния набор от необходими характеристики, гарантиращи качеството на данните при информационни системи, приложими в индустрията.

Ключови думи: качество на данните, сигурност, технологии, индустриални информационни системи

Introduction

Every individual and organisation needs the most current, accurate, and comprehensive information on the basis of which to take effective decisions.

This requirement is of particular importance for industrial information systems, where processes are in continuous dynamics and each of them can affect the performance of the entire system.

This is applicable to the greatest extent in the mining industry, since all processes are interrelated; they also depend on natural resources and require large investments in resources and tools (Eftimov, Anastasov, 2011). In this case, an inappropriate decision made on the basis of poor information may lead to huge losses for the particular enterprise.

In order to avoid such situations, it is highly important to obtain quality data, i.e. data conforming to the requirements of the specific information system. Data quality is directly related to the purposes for which they will be used (Tudjarov, 2012).

By definition, data quality is a characteristic that indicates the degree to which they can be analysed and to meet the needs of the business. From the point of view of information systems data quality is part of the whole process of data management.

Criteria defining data quality

The criteria according to which data quality is defined can be examined in two main aspects – from the point of view of their users and in terms of the possibilities for their usability in an information system (i.e., from the point of view of information technology).

Criteria for data quality, meeting the users' requirements

From the point of view of data users (i.e. people who take adequate decisions) the criteria for data quality can be considered in four large groups, namely availability, usability, intelligibility and security (Fig. 1).

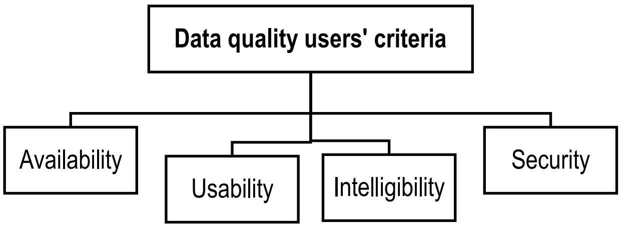

Fig. 1. Users' criteria for data quality

Availability of data

Availability of data means that in every moment, when appropriate, users need, to have access to them and they are always available.

In information systems, basic characteristics about the availability of data are accessibility, authentication, authorisation, and timeliness of equivalence.

In the client-server technology used by modern information systems the levels of access to a specific collection of data are defined at the design stage and an access level is assigned to every particular user, which determines what kind of data to be submitted. Various collections (databases) available for specific levels may exist. An example in this respect are the geographical information systems (Kazandjiev, Yanev, 2012), where there is different accuracy (data quality) depending on the type and level of access.

Depending on the specific level of access, it is verified if that user has permission (authentication) to use the information resource (i.e. to a lower or higher quality data). Authorisation is performed by the information system itself, as it gives the user rights to perform the permitted set of actions.

Since a large part of the information systems, including industrial ones, are used by many users, and different users can enter information, the equivalence of data is of particular importance. It measures the extent to which equality (equal values) of the same data is guaranteed.

The timeliness guarantees users that data are timely (as timely as possible), which is essential in making effective decisions.

Usability of data

The usability criterion means that data incoming in the information system from different sources can be processed and analysed. The data characteristics that determine their usability are documentation, validity, applicability, precision, flexibility and interactivity.

The most important feature of usability of incoming data is their ability to be converted into a digital format by the information system, i.e. they can be formalised by meeting their set conservation model (Kutzarov et al., 2012).

The validity of the data is determined by comparing the relevance to the requirements set for the specific information system.

Applicability is a characteristic that determines how much data can be processed and analysed in support of specific targets. In order to have adequate solutions taken on the basis of the data it is necessary to have precise data – i.e. they need to have values in the range specified in the information system. Thus, the level of detail of the data, which is required by different groups of users and management levels, is defined. The too high level of refinement and detail of data often leads to difficulties in the operation of information systems and it is therefore necessary to find a level of balance that satisfies both characteristics at one and the same time.

For the data to be used by different management levels (different user groups) and to be available on different devices (PC, Tablet, Smartphone), it is necessary to possess flexibility, which is particularly important in ERP systems. This means that they are subject to processes for different organisational changes or reengineering with minimal modification of the existing objects and relations in them. The use of information systems through the Internet or in a network mode requires the data to be interactive – that is, to have two-way communication between the data and users.

Data security

Data security assures the users that they are provided with the requested information in an accessible form and the data origin is guaranteed. The main features ensuring data security are standardisation, reliability, comprehensiveness, integrity, objectivity, comparability and stability.

Standardisation ensures that the data submitted and processed correspond to the rules set in each information system, which in some cases are valid for different information systems that share and exchange information. This data feature is set in the design process of the relevant information system and is monitored throughout its entire life cycle.

Nowadays, the reliability of data is a key feature not only for information systems but also for society as a whole. They give confidence about the source of the data and its reputation, which determines the degree of confidence in the data. Comprehensiveness is a complementary feature that determines to what extent the data is satisfactory and covers the user's request. Data integrity is one of the most important features of data, especially in an insecure environment such as the Internet, because it ensures that changes to data are made only by authorised users. The objectivity feature of the data ensures that the data are not modified under the influence of human emotions, i.e. only the specific facts about the data are reflected.

On-demand information systems should allow for comparability of data, i.e. to check that their values are the same in different systems (Arsova, Hristov, 2018).

Naturally, one of the most important features of data is the ability to be permanently stored and accessible over a long period of time to ensure its stability.

Criteria for the quality of data meeting the requirements of information technology

The requirement to use high-quality data in information systems, on the basis of which correct and effective solutions are available (nekdata.com), must meet at least five basic criteria − completeness, accuracy, validity, consistency and timeliness (Fig. 2).

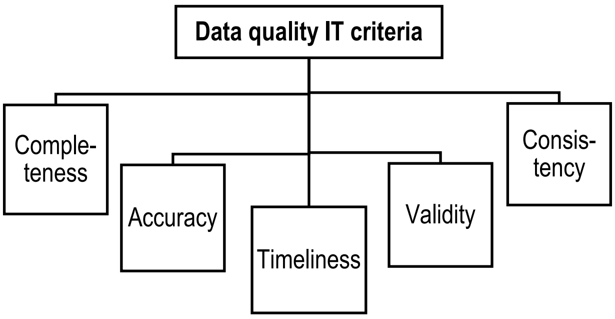

Fig. 2. Main criteria for data quality

Completeness of data

Unlike standard data collection (on paper), information technologies make it possible to ensure the completeness of data by using functions that allow the input and digital storage of information only where all attributes for the object, activity etc. have been introduced.

To ensure full quality data, additional features are introduced that check not only the correctness of the data provided but also the exact implementation of the data entry format defined by the particular information system.

Accuracy of data

Accuracy of data criterion suggests that incoming data in the information system are correct and fully reflect the depicted object, process, etc. To avoid the risk of inaccurate data submission, the interference of the human factor in this activity should be minimised already at the design stage of a specific information system. Unfortunately, this is almost impossible, and therefore, the implementation of this activity must be done by competent and well trained specialists.

To ensure the data accuracy, especially in cases of a high volume or a continuous stream of data, additional features are being set in the information systems which check for inaccuracies at every step and eliminate admission of such.

Data validation

The criterion validity of the data determines how data values are correctly measured according to the pre-set conditions. If we have received invalid data, this means that there is a problem in the process of collecting the data.

When you get values for specific data that are beyond the limits of the usual, it does not always mean that they are invalid. In such a case the values should be re-checked. In the flexible information systems this problem is easily solved by altering the defined limits for measured values and incorporating new values.

Consistency of the data

In information systems, especially in those with longer term of use, there are data about the same object, process, action, etc., that are introduced at certain periods and have different values. In other words, there are different versions of the data for an object or process.

The consistency criterion ensures that the data in the various versions are saved in the same format and most important, this data format is not changed during processing.

Timeliness of data

In order for an adequate and efficient decision to be made, it is important that the data we need to analyse should be timely − i.e. there is no time interruption of the incoming data stream for various reasons.

The timeliness criterion is especially important in industrial systems, which manage continuous production processes because the lack of data for a specific segment of time can lead to incorrect management decisions.

Data quality assurance technologies

Information technologies use a variety of techniques to ensure high-quality data needed to make effective management decisions. Out of this group, we can distinguish as particularly critical for the quality the technologies that provide standardisation, profiling, matching, control and clearing of data in real time (Fig. 3).

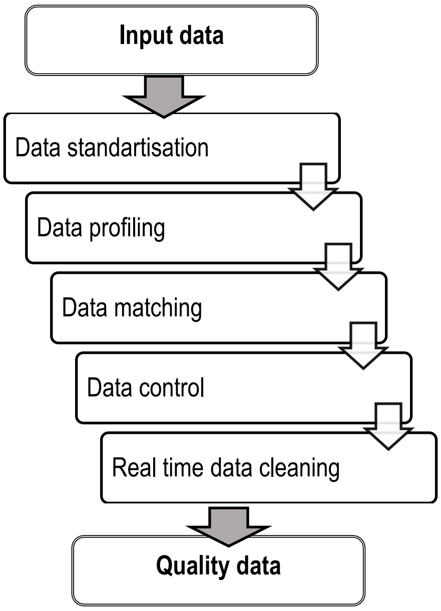

Fig. 3. Data quality assurance technologies

Data standardisation

The standardisation process is generally the affixing of various variables on the same scale. This process allows comparison of the results obtained from different types of variables.

In information systems, data standardisation is a tool that acts on the basis of set rules and ensures that the data comply with the specified quality criteria.

Data standardisation is the critical data input process in the so-called common format for large information systems used over the Internet. In order to ensure the quality of the data received, they undergo different transformation processes to meet the rules laid down in the specific information system (Yanev, 2013). Furthermore, in the business logic of these information systems, additional functions are provided to allow automatic correction of minimal inaccuracies and rejection of data in case of significant discrepancies.

The standardisation is of particular importance in ERP systems – information systems that allow management of all business processes in a big company, where the information comes from different sources. This technology is essential also when we have exchange of data between different information systems.

Data profiling

Data profiling is a technique used to analyse the content, quality, and structure of output data, and is used in various criteria for data quality, such as determining their accuracy and completeness.

The data profile contains the definitions of the sources, functions, and functional parameters and the parameters of the profile session. This process examines the data sources by initially evaluating the data to identify potential and actual shortcomings. The goal is to find out the wrong areas in the data organisation that can be found in user input, interface errors, data corruption when transferring, and so on. The use of this technique significantly improves data quality.

Data matching

Data matching is a technique for finding records that relate to the same object, process, individual, etc. Typically, these records come from multiple datasets and do not have common object identifiers, but data matching techniques can also be used to detect duplicate records in a single database.

In information technology, data matching can be done in many different ways, but the process is often based on algorithms or programmed circuits, where processors perform sequential analyses of each set of data by comparing it with each separate part of another dataset or by comparing complex variables to find strings containing specific resemblances.

Data matching establishes links between similar, yet different, records using the set data matching rules.

Data control

Data control is a set of techniques that monitor for changes in data quality over time and notify about deviations in the pre-set quality indicators. Data control can be implemented through various technologies in modern information systems.

The completeness of the data as a quality criterion is realised through the so-called "mandatory fields" that do not allow incomplete data to be received.

The data accuracy and validity can be realised using the so-called "drop-down menus", where a value can be chosen only from the ones defined in the system.

The timeliness of data in information technology is most easily ensured through cloud structures where all data about a particular object, process, individual, are automatically transferred to the cloud once the process completes and become immediately available to all users authorised to work with them.

Real time data cleaning

Data cleaning is a process of identifying incomplete, incorrect and inaccurate data. The clean-up corrects or removes damaged or inaccurate records as well as inappropriate sections of data, and then replaces, modifies or deletes the so-called contaminated data.

This is the process that ensures that the data is correct, consistent and applicable. Data clearing is important because it improves data quality by removing any obsolete or incorrect data and leaves the highest quality information.

Information technologies allow data cleaning and quality control processes to be embedded in the relevant applications in order to be implemented in real time. This in practice does not allow input of incomplete, inaccurate and invalid data.

Conclusion

The data quality is of particular importance for all modern information systems that operate in almost all areas. It is important both for the business as a whole and for a particular process, action, individual, etc.

The quality of data is of paramount importance in making informed, adequate and effective decisions, especially in the areas of national security, which include the mining and energy sector. Many big mining companies plan, control and manage their operations through specialised information systems tailored to their specific needs. Each mining company, depending on its specificity, determines which data is essential for management decisions, i.e. it defines its own set of high-quality data, the ultimate goal being to get quality product at optimal cost.

As can be seen from the above-mentioned, the criteria for data quality from the users' point of view do not fully match the techniques that guarantee the data quality via the information technology. The implementation of all of these criteria at the same time is not an easy task. It is therefore necessary to find the right balance between the consumers' requirements, the information technology and the most relevant criteria for high-quality data on a case-by-case basis.

References

Arsova-Borisova, K., V. Christov. 2018. Models in designing document management systems. − *Journal of Mining and Geological Science, 61*, 3, 70−74.

Eftimov, Z., D. Anastasov. 2011. Scientific Aspects in Formation of Quality of Ore in Extraction Stage. − *22nd World Mining Congress*, *I*, 181−185.

Kazandjiev, I., N. Yanev, K. Ivanov. 2012. Graphical representation of data with open source software. – *Annual of UMG "St. Ivan Rilski", 55*, 4, 123−127 (in Bulgarian with English abstract).

Kutzarov, K., D. Anastasov, Z. Eftimov. 2012. Principles for the application of mining software for the planning of mining processes in the extraction of underground natural resources. − *Journal Mining and Geology*, *2-3*, 56−58 (in Bulgarian with English abstract).

Tudjarov, H. 2013. Data Management, Asenevtsi; *http://www.tuj.asenevtsi.com/Data/IndexD.htm (in Bulgarian)*

Yanev, N. 2013. *Methodologies and technologies for development of information systems.* P*ublishing House "St. Ivan Rilski"*, Sofia (in Bulgarian with English abstract).

https://nektardata.com