# On Quality of Measurement Data in the Process of Knowledge Acquisition

Adam Markowski[1], Wiesław Miczulski[1], Robert Szulim[1]

[1] *University of Zielona Góra, Department of Electrical Metrology, Podgórna 50, 65-246 Zielona Góra, Polska, phone +48683282234, fax +48683254615, {A.Markowski; W.Miczulslki; R.Szulim}@ime.uz.zgora.pl*

***Abstract-*** In the paper, influence of quality of measurement data on quality of data warehouse, which is required in the process of knowledge acquisition, has been shown. Factors crucial to quality of data in the data warehouse, criteria of quality of data warehouse and indices of quality of measurement data have been presented. These elements indicate requirements for measurement systems and processes of transforming data from sources into data warehouses.

## I. Introduction

Proper conducting of the complex technological processes imposes the necessity of getting a chain of different information from analyzed objects. The quality of processing that information is related to complexity of the diagnosed objects and form of description of the rules of their functionality. Analytical methods have limited usability, because of problems with getting accurate models, which are usually not linear. Alternative solution is using qualitative models based on techniques and methods of intelligent computations. Expert systems, artificial neural networks and fuzzy sets theory can be used as this kind of techniques. From methodological point of view, using expert systems to aid conducting of technological processes is an attractive solution.
The knowledge base, which has an influence on quality of elaborated expert system, is very important issue to solve. The process of creating the knowledge base is connected with its acquiring from experts and from databases containing results of the measurement of the values describing technological process.
Present experience of the authors indicates that for very complex technological processes (like metallurgical processes) quality of created knowledge base depends on selection of the experts and way of their acquisition [1]. In the group of the experts, it is possible to mark direct and indirect experts. Direct experts take personal part in the process of knowledge acquisition. There are operators of the process, group leaders, and managers, supervisors, who are responsible for getting rules, explanations, semantics, and procedures of the process. They are also responsible for creation and maintenance of the dictionaries of conception. This stage of knowledge acquisition requires the usage of special forms to allow building of the knowledge base in a hierarchical way. The knowledge acquired in that way often has subjective character. Indirect experts are authors of different publications (documentation of the object, description of the process), which are respectively interpreted by third persons. The knowledge base, which is being created, may contain cohesion and redundancy errors, which might be difficult to discover at its process of logical verification stage.

Fast development of the informational techniques and tools is being observed in the few past years. Using measurement data in the process of knowledge acquisition about objects and technological processes is possible with those tools. The knowledge being acquired from those sources has subjective character.

### A. Measurement systems as a source of the data

Measurement-control systems are very important in the process of the supervision of the state and control of the technological processes (figure 1). Amount and quality of information, which is being delivered to the operator of the technological process, may often cause problems with its proper interpretation. There are special computer and software systems of the SCADA class (*Supervisory Control and Data Acquisition*), which can aid operators in the process of supervision and control of the given object and technological process. The quantity characterized state of the object and process after preprocessing by intelligent measurement converters into digital data are processed in system controller

according to strict algorithms and presented on so called synoptic screens on computer monitors. The SCADA systems can aid operators in making fast and right decisions by supervising values of measured quantities, generating alarms after exceeding the range of the measured quantities, presenting trends, etc. Unfortunately, systems of that class usually are not equipped with elements of artificial intelligence, which means, that they can not analyze results of the measurement and take or suggest decisions.

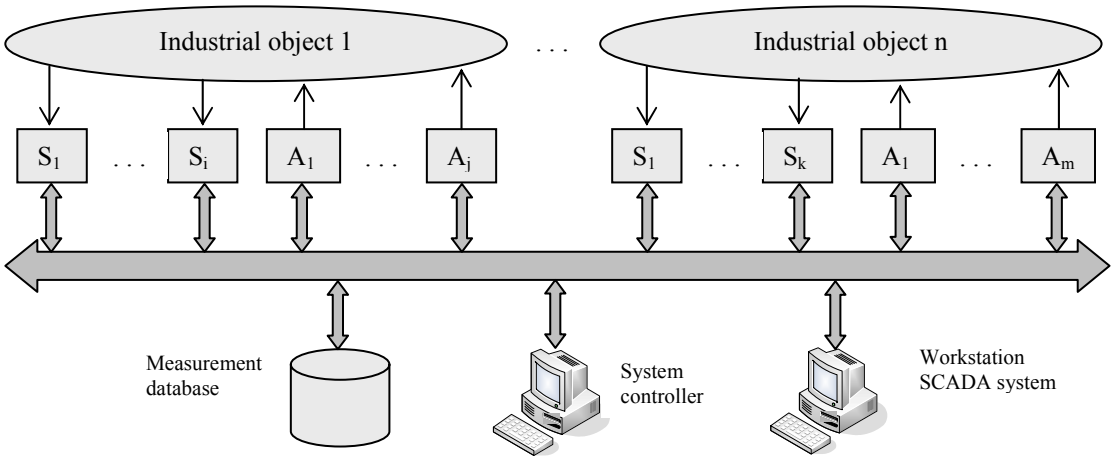


Figure 1. Measurement-control system; S – sensor, A – actor

Measurement-control systems beside direct influence on the object or process may also be the source of the data in the process of knowledge discovery of objects or technological processes.

**B. Data warehouses**

The process of the knowledge acquisition might be defined as group of techniques of automatic discovery of non trivial and unknown relations and schemas (data mining) in large data sets. Those large datasets create analytical database called data warehouse, which is a foundation of the decision support system. The structure of the warehouse consist of three basis data layers (figure 2): data sources, central warehouse and topical warehouses. The data of one layer are derived from lower layers.

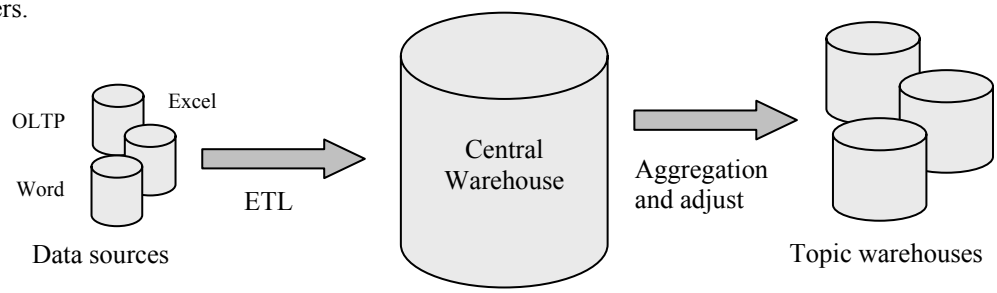


Figure 2. General architecture of the data warehouse

Depending on the profile of the company data warehouses base on sources containing data important for certain departments, such as marketing, sales, finance. As separate group of data making source of the data for data warehouse, measurement data from technological processes or observations of technological object might be named. SCADA systems are used in the process of acquisition of the data. Production companies use both of the data group during the process of building the data warehouse. The process of collecting and storing of the data is not fully reliable. The number of incorrect data in the database in different application can reach in practice 10% or more [2]. Thus, very important problem is the quality of the data.

## II. Quality of the measurement data and quality of the data warehouse problem

Quality of the measurement data in the data warehouse depends on: the project of the schema of the data warehouse, quality of the data inserted into data warehouse, data processing in the data warehouse. Each of the mentioned factors depends on set of tools used in process of creation data warehouse. The

quality of measurement data also depends on features of measurement systems, which are source of that data.
The process of elaborating the schema of the data warehouse may have complex character. It covers requirements analyzes, available data analyzes, source schema acquisition and their integration and also other activities connected with process of projecting databases. Proper schema of the data warehouse decides about suitable, complete and significant integration of the data sources. Improper interpretation of the semantics of the source data or improper integration of the data sources may cause the situation, when data warehouse will be filled with improper information.
The most important matter is the quality of the source data used during the process of creation and actualization of the data warehouse. Quality of source data decides of the quality of the data warehouse. Data with errors may me moved into data warehouse. Mechanisms of loading and actualization may also have influence on data inserted into data warehouse.
Processing of the data in the data warehouse is connected with aggregation and multidimensional data reorganization. It means that quality of the data inside the data warehouse is in general proper and processing of the data has insignificant influence on the quality of the data.

**A. Data warehouse projecting**

The process of design of the data warehouse should begin from building the model of the company, which also connects semantic models to each other in the source layer, covering different data sources (Figure 2). This approach to project of the warehouse is expensive and time consuming, but can be a basis for further development of the schema of the data warehouse. The central data warehouse schema should contain two primary layers. The first layer consists of operational data store (ODS). This layer is responsible for cleaning and transforming the data. The second layer covering unified data is actual analytical database.
Designed and built data warehouse should meet following quality criteria [2]:

- Accessibility – how to facilitate the better access to data for the user.
- Usefulness – how to access to the data warehouse will meet requirements and work conditions of the user.
- Reliability – how to convince the user about reliability of the data and how to improve reliability of the data acquired from not reliable sources.
- Interpretability – how to facilitate understanding of the data for the user.
- Verification – how to check if all above factors were considered.

**B. Measurement data quality**

The three first criteria of quality of the data warehouse directly refer to data sources containing also measurement data. From the measurement data point of view, reliability criterion is important, which is directly connected with following measurement data quality indicators: precision, completeness, actuality and cohesion.
Precision and completeness of measurement data are determined by the project of measurement system phase. In classic measurement system (Figure 1), we often have to deal with a fixed structure of the system. Introducing a new element to the system demands modification of the code of application in every element of measurement-control system, which would operate on data coming from that new element. It complicates significantly phase of implementation of the application in a given element, complexity of the algorithm thus probability of existence of errors in the application increases. This situation may influence in a significant way the stability and proper work of that element or even whole system. This also means that the occurrence of incorrect data values or their loss is possible. Using XML techniques during the process of projecting measurement-control systems can help to avoid aforementioned problems, which result in increasing precision and completeness of the data [3]. The loss of measurement data may also occur in case of improperly chosen values of parameters vital for communication features, ex bus transfer rate, algorithms of chaining [4]. Precision of the measurement data may also be affected by metrological, static and dynamic features of instruments used to build measurement systems. More frequent usage of intelligent converters in measurement systems, with algorithms of static and dynamic errors correction help in meeting the criteria of keeping imposed precision of values of the measurement. Those issues are broadly presented in the literature. In order to improve quality of the knowledge from measurement data, it would be advisable to store results of the measurement together with its uncertainty. Calculation procedures realized in intelligent measurement converters must be supplement with algorithms of calculation of uncertainty of values of the measurement for non stationary values [5]. Reliability of measurement systems work may also have an impact on precision and completeness of the data in data sources. In [6] there are results of analyses of simple diagnostic systems, which state that measurement systems may be characterized by even 40 % of failures. It means that there is a need of an automatic diagnosis of the elements of measurement

systems to check reliability of their work.
Much bigger problem connected with keeping accuracy indicators can be unreliable manual records in data sources, ex in Excel sheet, results of measurement outside of the SCADA systems (ex, chemical analyses results, which are conducted in laboratories outside of the measurement systems [7]). These data might have improper values. Lack of manually entered values can also happen, ex in Excel sheet. It means that quality indicator, which is completeness, is not met.
Actuality is an indicator of quality of data (not only measurement), stored in data warehouses informing, if data stored are not out of date. This indicator influences planning strategy of updating the data in a data warehouse. On the other hand, the cohesion specifies the uniformity of the data according to none conflicting of information stored in a data warehouse.
Improvement of the quality of the data inserted into warehouse can be achieved using ETL programs (*extraction-transformation-load*), which can do following tasks:

- extraction (access do different data sources),
- cleaning (detection and solving lacks of cohesion of source data),
- transformation (ex. between different formats of the data, different languages),
- loading (creating copy of source data in the data warehouse),
- analyses (ex. detection improper or not allowed values in data ),
- fast data transfer (important in case of using huge data warehouses),
- quality of data inspection (ex. by means of completeness and correctness),
- metadata analyses (by means using them in the process of projecting data warehouses).

Improvement of the quality of data with using ETL programs results in decreasing of amount of the data stored in warehouse. Technological processes are characterized by some periodicity. Loss of even small amount of measurement data or problems with incorrect data in one technological result in erasing all of the data connected with this cycle from data sources. In consequence, even 50% of the overall records of data stored in warehouse might be lost or impossible to use [7].

## III. Conclusions

Quality of the measurement data stored in data warehouses during the process of their creation and updating decide on quality of the warehouse and in consequence on quality of knowledge acquired from that data. Providing high quality of the warehouse impose certain requirements to meet for measurement systems together with loading and updating mechanisms. The high quality of the measurement data is connected mainly with the process of projecting of the control-measurement systems.

## References


[1] Bolikowski J., Sarafin M., Michta E., Miczulski W., Szulim R., „Export system to aid conducting process of slag reduction", *IV National Science Conference – Knowledge engineering and export systems,* Wroclaw University of Technology Press, Wroclaw 2000, T. 2, pp. 229 – 236.
[2] Jarke M., Lenzerni M., Vassiliou Y., Vassiliadis P., Data warehouses - basis and functionality, *Wydawnictwa Szkolne i Pedagogiczne S.A*., Warszawa 2003.
[3] Michta E., „Internet techniques in distributed measurement-control systems", *Computer aided metrology – conference materials*, t 2, pp. 223 – 232.
[4] Markowski A., „Estimation of the data transmission delay in network measurement-control systems", *Pomiary Automatyka Robotyka*, no. 7-8, pp. 95 – 99.
[5] Domańska A., "Aspects of uncertainty estimation of non stationary values", *Problems of the metrology – IV conference materials*, Work of the PAN Metrology Commission Department in Katowice, Katowice 2000, pp. 313 – 320.
[6] Isermann R., Balle P., "Trends in the application of model-based fault detection and diagnosis of technical processes", *Control Engineering Practice*, vol. 5, no. 5, pp. 709 – 719, 1997.
[7] Szulim R, "A method of mining knowledge to aid a control of complex industrial processes", *PhD dissertation*, Zielona Góra 2004.