Towards a Data Quality Policy for Supervision Data in CERN's Technical Control Room

R.Martini, P.Ninin
Robin.Martini@cern.ch
European Organisation for Nuclear Research - CERN,
1211 - Geneva - 23, Switzerland

Abstract

The implementation of the Technical Data Server (TDS), has highlighted the reliance of many of the software products used for the supervision of the technical service infrastructure at CERN, on the quality of the data used to represent the different processes being monitored. Many of the modules that make up the TDS, as well as many of its client applications, are data driven, and obtain their data from a relational database designed for this purpose. It is critical for the correct operation of this software, and hence for the quality of service provided to those involved in the supervision of the technical infrastructure, that the data used is up to date, correct and meaningful. The paper focuses on the acquisition phase of the data life cycle as this is where one can ensure that only the necessary and correct data are captured onto the database. A description of the tasks that need to be accomplished and the issues of responsibility are examined with a view to establishing procedures that encourage data quality.

1 Introduction

This paper addresses issues concerning data quality in the computerised surveillance of CERN's technical infrastructure which is under the responsibility of the Technical Control Room (TCR). The reasons for adopting a data quality policy as well as the changes necessary to its implementation are discussed.

2 Background

In order to keep CERN's particle accelerators operational, a large number of technical services must be provided. These are monitored centrally by the TCR which has evolved in just a few years from the surveillance of under 5000 data points (alarms, measurements, states and commands) using a stand alone “off the shelf” industrial package, to a heterogeneous system incorporating software developed by separate teams at CERN and a range of commercial systems, monitoring up to 100,000 points, covering a more varied and complex spectrum of equipment than before. Furthermore, the staff involved in the surveillance activities has become more volatile, so that whereas in the past, knowledge could be accumulated through many years experience, much information such as precise locations of equipment at the source of a problem, implications of a fault, people to call, precautions to take etc. must now be assimilated in a short space of time by operators who are in service for an average of 4 to 5 years. In order to manage this situation, the Technical Data Server (TDS) project was launched with a view to standardising and improving the control system software. The increasing complexity of this software prompted the introduction of techniques to improve software quality, however, since the TDS will ultimately be integrating data supplied by the different users of this software, a commensurate attempt should be made to ensure the quality of the data it uses.

Databases have long been used to store key enterprise data at CERN, but each application area tends to build and populate its own database without concern for others which hold or require related data.

3 The case for a data quality policy

The following issues highlight the importance of data in the TCR monitoring environment : the "data driven" nature of the monitoring software, the value attached to data by the users, the potential information that can be drawn out of the data and the feeling towards the current quality levels.

3.1 A data driven monitoring system

The TDS was designed as a data driven system in order to cope with the heterogeneous control environment; therefore the quality of the data that is handled by the TDS impacts directly on the quality of the services it provides. Furthermore all client applications of the TDS, the alarm system, the mimic diagrams and data logging system rely on this data to perform their functions. All data that is used by the TDS is held in a relational database (TDRefDB); this data is called the static data (as opposed to the dynamic data which consists of the data's real time value.), it allows operators and other users to understand the behaviour of the systems being monitored and may indicate what action to take in certain situations.

3.2 Data quality impacts operation

Should data that operators be dealing with be incorrect, ambiguous or otherwise unreliable, confusion may arise in communicating the state of a system being monitored to other services at CERN (users, maintenance personnel etc.); incorrect actions may then be taken with severe consequences, this undermines the trust that the operators (and other users) place in the monitoring system and can adversely affect both motivation and job satisfaction.

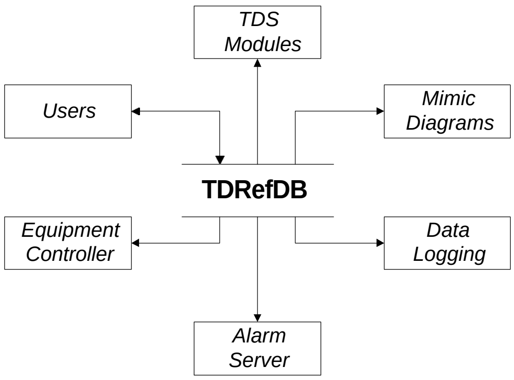

Figure 1: Role of the TDRefDB

3.3 Strategic information availability

Due to the changing requirements of the organisation and to the need to renew older equipment, the systems being monitored are constantly evolving : it is important to be able to know accurately and quickly the details of any system being monitored, such as the number of and types of alarms and commands being managed, so change can be planned and implemented satisfactorily. One of the aims of the TDRefDB is to provide users with such "documentation" on what is being monitored, but again a lack of quality in the data renders this documentation of doubtful use.

3.4 Current quality levels

Though there has been no systematic study to assess the current level of data quality, anecdotal evidence from operators and other users, reporting unclear messages or identifiers that do not match real objects, points to the need to address this issue. Inconsistent naming also hampers the operators' understanding of the processes they are monitoring, since they must deal with a wide variety of systems that may each have a different way of describing their environment.

4 Scope of the policy

The life cycle of the static data in the TDRefDB can be broken down into three separate activities: (1) acquisition, (2) storage, and (3) usage. The acquisition activities are concerned with all processes where a human agent can affect the values of the data to be stored; this includes data destruction, which is considered as the assignment of the property "invalid" or "out of date" to a data structure. The storage activity is handled by a centralised relational database, regularly backed up and is considered as providing a high quality service. The usage activities are determined by the programs that treat the data they require from this database, the quality of the data output is determined by the quality of those programs as well as that of the input data.

It is suggested here that the policy should focus initially on quality improvements of the acquisition cycle as the most significant gains to quality can be achieved in this domain, the storage and usage activities are not treated here, though should form part of an eventual data quality policy.

5 Data characteristics

The data on the TDRefDB has a fairly low volatility and most modifications are due to corrections of inaccuracies. Major changes to data occur when new systems are taken on, requiring the input of large volumes of data to the database. Similarly when systems or equipment are decommissioned, data must be removed from the database. A brief description of the different types of data treated and the category of users concerned by them is given.

5.1 System data

This covers data essential for the configuration of the software systems involved in the monitoring processes, such as surveillance program names, host machine names, and some domains such as point types or alarm levels. This data is supplied by and is under the responsibility of the TDS administrator.

5.2 Equipment data

This covers all data that describes the points being monitored such as the description, location, classification codes, person responsible; the data supplied differs depending upon whether it concerns an alarm, state, measurement or command, though certain parameters apply to all points on the database. The equipment specialists have the responsibility of the upkeep of this data, though they may need to obtain some of it from the suppliers of the equipment level control system.

5.3 Operational data

This concerns data that is used in the monitoring activities and includes alarm help information and the parameters of the different mimic diagrams, such as the list of points that are used by each one. The operators are responsible for the upkeep of this data, but rely on the equipment specialists for knowing what to supervise.

6 The data acquisition life cycle

The first systems to be managed by the TDS showed that a systematic approach to the problem of populating the database with the appropriate and correct monitoring data had to be adopted. To achieve this, the acquisition activities of the data life cycle are subdivided into five phases: definition, preparation, capture, maintenance, and destruction. The tasks necessary to ensure optimum data quality are detailed within each phase.

6.1 The data definition phase

This phase principally concerns the taking on by the TDS of a new monitoring application, or the major extension of an existing application; the output from this phase will be:

(a) the list of points (alarms, states, measures and commands) that will be used in the new monitoring application,

(b) the domains of the points in the application,

(c) the data necessary for the functioning of the TDS,

(d) the necessary database structure changes (if any),

(e) the monitoring tasks assigned to the TCR.

An analysis of the monitoring application must be carried out to obtain the above information. This process is under the responsibility of equipment specialists, and requires the collaboration of TDS and database administrators, the equipment layer control system specialist and a representative of the TCR.

For minor changes to existing applications, consisting typically of the inclusion of a small number of new points, the data to be entered need only be defined in collaboration with the operations section to ensure that their monitoring applications can be updated.

The database administrator is responsible for carrying out the changes to the database that have been identified during analysis, (this could be the case when the TDS is to communicate with a new type of equipment controller), so that the structure and format of the new data can be established. The TDS administrator may then identify the necessary changes to client applications that use this data.

6.2 The data preparation phase

Once the data necessary for monitoring has been identified and defined, the equipment specialists are responsible for assigning the appropriate parameters to each point (such as description, location, hardware address, data type, priority etc.). The TDS administrator is similarly responsible for allocating the new values to the system related data elements. The output from this phase is:

(a) the new TDS data to be entered,

(b) the complete list of new data points and monitoring parameters, agreed to by equipment specialists and operators.

When the volume of data to be input is large (more than 50 points), the TDS administrator may provide the equipment specialist with an Excel sheet that has been formatted to accept the relevant parameters. This can help speed up the transcription of data, as well as facilitate data verification by highlighting points that are missing or have inappropriate parameters. This document is used in the next phase to load the data directly onto the database.

The TCR responsible ensures that standard definitions are used wherever appropriate as this helps operators make sense of the heterogeneous monitoring environment. Any naming conflicts are identified at this stage, and resolved between the TDS administrator and the equipment specialist. The operations staff are consulted when assigning descriptions and codes (e.g. equipment codes) to points since these elements must be meaningful to all people who will be using the surveillance tools where they are displayed.

6.3 The data capture phase

This phase is the conclusion of the acquisition activities concerning new data, the output from this phase is:

(a) error reports indicating incorrect data,

(b) the data stored on the database,

(c) reports detailing the new data input.

The TDS administrator updates the database with all new monitoring system data; this task must be completed before any other data is entered since the equipment data may depend on it for validity.

Once the agreed list of new data points has been established, it can be input to the database. When dealing with large quantities of data, standard database loading programs based on the predefined Excel documents can be used. Integrity and domain constraints are built into the database and are maintained by the administrator; should these constraints detect any errors, the equipment specialist is informed so that corrections can be made; for a large number of errors, the corrected input document is reloaded after correction.

An HCI is provided to the equipment specialists for the input of small quantities of data. To ensure that only correct data is entered, all violations of integrity and domain constraints are rejected and reported to the user; help screens showing valid values are available where appropriate. The HCI is also designed to guide the user through the correct steps in data entry.

When all equipment data is successfully loaded onto the database the mimic diagram designers (often TCR operators) may specify the points that appear in these applications. Reports are available to the people concerned by the new data so that a final check can be made. An equipment specialist may obtain a list of all the points he is now responsible for, after addition of the points for the new application; both operators and equipment specialists have access to a list of points appearing on the HCIs, or of the alarms to be sent to the TCR.

6.4 The data maintenance phase

This phase covers the ad hoc changes that need to be made to the database, when changes in a monitored system have been made, errors have been spotted and supplementary information can be added.

The TDRefDB HCI (as used in the data capture phase) allows these changes to be made. The HCI must ensure that only the appropriate people can make changes to the data. The database administrator has the responsibility of ensuring that each user can only access their own data, for example each equipment specialist can only modify his own data, similarly only operators and the TDS administrator can modify mimic diagram information and monitoring configuration data respectively.

6.5 The data destruction phase

This phase concerns the removal of a point no longer monitored, or when a system or large part of a system is decommissioned.

The equipment specialist is responsible for determining which points can be removed from the TDRefDB and mayeither use the HCI to achieve this or supply the database administrator with a list of points to be removed if a large number is concerned. A flag is set against any point that is to be removed, this ensures that the point cannot be used by any part of the monitoring software, the operations staff will be informed of the points whose removal has been requested list so that the surveillance programs can be modified accordingly.

Once agreement between equipment specialists and operators has been reached, the database administrator definitively deletes the point details from the database.

7 Conclusions

In trying to coordinate the surveillance of a large number of systems managed by different organisational units, the TCR is faced with the problem of integrating the data it receives from these systems into an accurate and meaningful representation of the real world. This task is critical to the core activity of CERN, i.e. the running of large accelerators, and for this reason the issue of data quality is deemed to merit attention. Our experience has shown that satisfactory quality levels will only be obtained by following established and agreed procedures for acquiring data. The tasks undertaken within these procedures will only work if they are assigned to the appropriate persons, similarly the responsibility for the successful completion of each process must be correctly defined. For each new project, sufficient resources in time and personnel must be allocated to the data acquisition phase.

When large amounts of data are taken these tasks can be managed with GDPM[1] techniques, by setting milestones and allocating staff to roles within the different activities.

An effort also needs to be made in defining standards across equipment groups, this is a difficult and sensitive task as it concerns people in different organisational units, however the benefits to be gained, in easier implementation of monitoring applications, and better central supervision by the TDS should encourage management to tackle this issue.