Future Circular Collider

PUBLICATION

Accelerator Reliability Information System Data Quality Concept

Böhm, Peter (AIT) *et al.*

22 January 2019

The research leading to this document is part of the Future Circular Collider Study

The electronic version of this FCC Publication is available
on the CERN Document Server at the following URL :
<http://cds.cern.ch/record/2654602>

ARIES

Accelerator Research and Innovation for European Science and Society

Horizon 2020 Research Infrastructures GA n° 730871

DELIVERABLE REPORT

Data Quality Concept

DELIVERABLE: AIT-IO-2.1

Document identifier:	1.0
Due date of deliverable:	End of December 2018

ABSTRACT

One of the requirements for a future reliability and availability information system is to ensure adequate quality of the gathered data. This document presents the state-of-the art in reliability data quality estimation. The document describes what qualifiers and control processes should be used to assess and ensure the data quality in the Accelerator Reliability Information System (ARIS).

For more information on ARIES, its partners and contributors please see http://aries.web.cern.ch

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No 730871. ARIES began in May 2017 and will run for 4 years.

Authored by

Name	Partner
Petr Böhm Thomas Gruber Alexander Preinerstorfer Heinrich Humer	AIT
Arto Niemi Johannes Gutleber	CERN

TABLE OF CONTENTS

1 Introduction

This work is part of ARIES EU project task to study feasibility of reliability information sharing within the accelerator community. In practice, this means establishing an Accelerator Reliability Information System (ARIS) where users could upload and access the reliability data. This approach has been successfully used in industry and this experience is detailed by our earlier literature review [1]. We have also published a report describing the use cases of the ARIS [2]. This report presents a literature review on maintenance data quality estimators.

Data quality is a context dependent term and the correct measure depends on the use case. Two main use cases can be identified: 1) Data is uploaded to the ARIS and the quality measurement shows if the data are technically correct to be uploaded. 2) An end user tries to understand how credible a data sample is for a reliability analysis. The current plan is that the ARIS would present anonymized information and a user would not know the source of the data. In this situation, a data quality measurement is an important indicator that describes how credible an individual data sample is.

This document presents the state-of-the art in reliability data quality estimation. The document describes what qualifiers and processes should be used to assess and ensure the data quality in the ARIS.

2 Definitions and State of the Art

In this section, some definitions related to the Data Quality and Data Quality Metrics and State of the Art will be described.

2.1 Data Quality Definition

Data quality is a perception or an assessment of data's fitness to serve its purpose in a given context.

The standard ISO-14224-2016 [3] claims, that the high-quality data are characterized by the following:
a) completeness of data in relation to specification;
b) compliance with definitions of reliability parameters, data types and formats;
c) accurate input, transfer, handling and storage of data (manually or electronic);
d) sufficient population and adequate surveillance period to give statistical confidence;
e) relevance of the data to the need of the users.

Data quality is a critical issue that should be considered – starting with initial application design, all the way through implementation, maintenance and use.

2.2 Data Quality Measures

The naming of different measures of data quality varies in the literature [4]-[8], here are the most common dimensions of data quality:

2.2.1 Completeness

Completeness is defined as expected comprehensiveness. Data can be complete even if optional data is missing. As long as the data meets the expectations then the data is considered complete.

(Is all the requisite information available? Do any data values have missing elements? Or are they in an unusable state?)

2.2.2 Consistency

Consistency means data across all systems reflects the same information and are in synch with each other across the enterprise.

(Are data values the same across the data sets? Are there any distinct occurrences of the same data instances that provide conflicting information?)

2.2.3 Conformity

Conformity means the data is following the set of standard data definitions like data type, size, range and format.

(Do data values comply with the specified formats? If so, do all the data values comply with those formats?

2.2.4 Accuracy

Accuracy is the degree to which data correctly reflects the real world object OR an event being described.

(Do data objects accurately represent the “real world” values they are expected to model? Are there incorrect spellings of product or person names, addresses, and even untimely or not current data?)

2.2.5 Integrity

Integrity means validity of data across the relationships and ensures that all data in a database can be traced and connected to other data.

The inability to link related records together may actually introduce duplication across your systems.

(Is there are any data missing important relationship linkages?)

2.2.6 Timeliness

Timeliness references whether information is available when it is expected and needed. Timeliness of data is very important.

The timeliness depends on user expectation. Online availability of data could be required for room allocation system in hospitality, but nightly data could be perfectly acceptable for a billing system.

2.2.7 Uniqueness

Uniqueness means nothing will be recorded more than once based upon how that thing is identified. It is the inverse of an assessment of the level of duplication

2.3 DATA QUALITY METRICS

The Figure 1 shows an example of Data Quality Metrics.

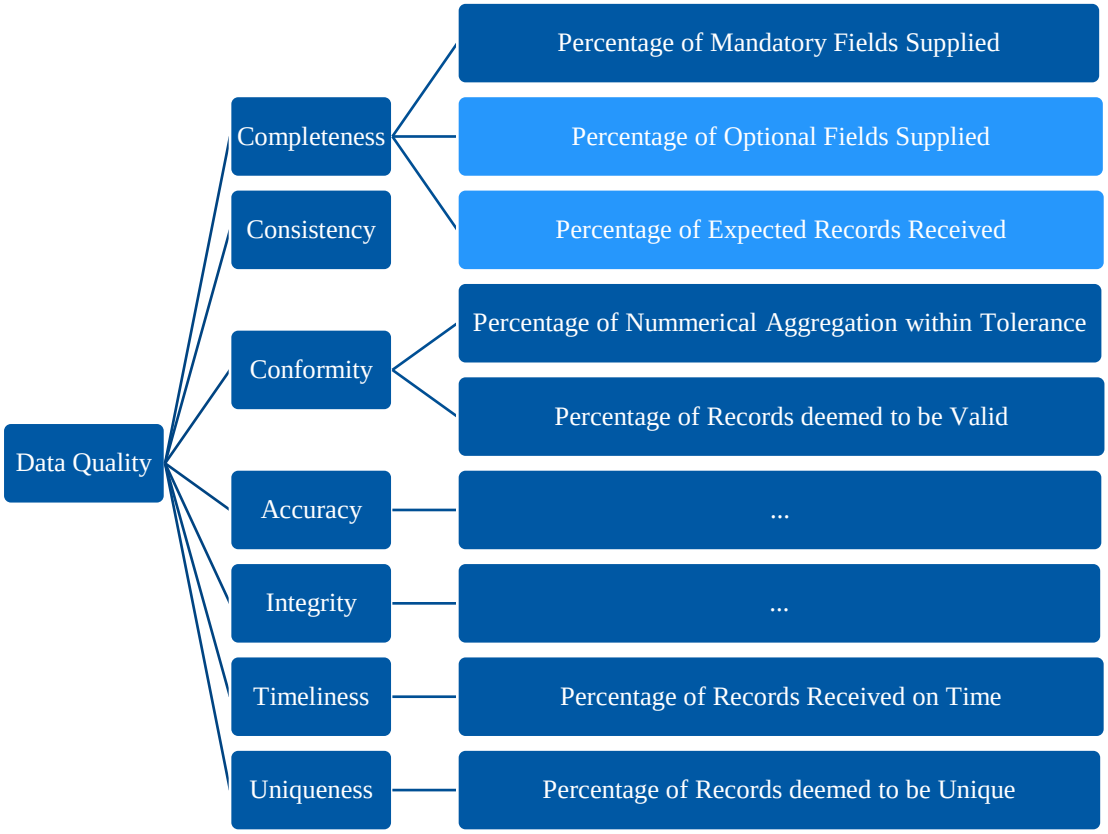

Figure 1: Example Data Quality Metrics

2.4 DATA QUALITY LEVEL

Ways to assess the quality of maintenance data have been studied in literature. Reference [5] presents a concrete schema to derive a data quality level by grading the detail and machine readability of different data fields. Alternatively, references [6] and [7] present a quality estimation approach that is based on analytical hierarchy process (AHP). This approach measures the quality of a reliability data sample in relation to other reliability data samples.

A concrete data quality level based on defined metrics would be more ideal than a relative metric. A quality estimate given by a relative metric depends on the quality of the data that is in the database. This dependency is a shortcoming of the AHP approach. If a data quality measurement approach is implemented, the approach presented in [5] could act as a basis.

However, all of the references assume access to the full miniatous data. A plan is that ARIS will use similar approach to the Fusion reliability database [9] and ISO 6527 [10] where only the failure rate and availability are stored in to the reliability information system. In this case, data quality level assessment cannot be implemented in the ARIS and the quality level should be assessed by data providers.

2.5 DATA QUALITY QUALIFIERS

In this section, some qualifiers which contain information on data quality will be described

2.5.1 Sample size

Number of units for which data are collected

2.5.2 Number of installations

Total number of installations (platforms) covered by the data surveillance for the item in question [11]

2.5.3 Length of the observation period

Total time of the observation period

2.5.4 Type of failure rate

Failure rate is

a) observed during operations
b) taken from literature
c) expert opinion

2.6 DATA QUALITY CONTROL PROCEDURE

The database should have some quality control procedure for the inputet data, so, that the errors in the data would be noted.

A few basic automatic checks could be carried out on all data. These include i.e. date and time (year 4 digits, month between 1 and 12, day in range expected for month, hour between 0 and 23, minute between 0 and 59), and range checks (tests that observed parameter values are within the expected extremes).

In addition, there should also be a manual verification procedure and the data should at least have qualifiers *not reviewed / reviewed*. Or, a detailed quality flag can be assigned to each data value, as shown in the Table 1:

Key	Data Quality Flag	Description
0	No quality control	No quality control procedures have been applied to the data value. This is the initial status for all data values.
1	Good value	Good quality data value that is not part of any identified malfunction and has been verified as consistent with real phenomena during the quality control process.
2	Probably good value	Data value that is probably consistent with real phenomena, but this is unconfirmed or data value forming part of a malfunction that is considered too small to affect the overall quality of the data object of which it is a part.
3	Probably bad value	Data value recognised as unusual during quality control that forms part of a feature that is probably inconsistent with real phenomena.
4	Bad value	An obviously erroneous data value.
5	Missing value	The data value is missing.

Table 1: Data Quality Flags

3 Conclusion

Chapter 2 provided an overview on reliability data quality concepts described in the literature. The methods for assessing data quality require access to the full dataset. The plan is that the ARIS would use a ISO 6527 type approach [10] where only the reliability metrics are stored in the database. If a data quality metric is implemented, a data provider need to be able to assess data quality independently. Here the method defined in [5] could act as a basis.

OREDA handbooks [11] and ISO 6527 [10] do not present data quality. Instead, they use data qualifiers that can be used to assess the credibility of data. These are presented in the section 2.5 and they should be available in the data stored in the ARIS.

4 References

1 A. Preinerstorfer, H. Humer, P. Böhm, T. Gruber, A. Noemi and J. Gutleber, Bibliography and state of the art of reliability information systems, ARIES note, (Zenodo, 2018). https://doi.org/10.5281/zenodo.1292068

2 A. Preinerstorfer, H. Humer, P. Böhm, T. Gruber, A. Niemi and J. Gutleber, Information system use cases and use context of an open reliability information system. ARIES note, (Zenodo, 2018). http://doi.org/10.5281/zenodo.1317768

3 International Organization for Standardization, Petroleum, petrochemical and natural gas industries - Collection and exchange of reliability and maintenance data for equipment, International standard ISO 14224:2016, 2016.

4 M. Hodkiewicz, P. Kelly, J. Sikorska and L. Gouws, A Framework to Assess Data Quality for Reliability Variables, in Engineering Asset Management - Proceedings of the 1st World Congress of Asset Management, pp. 137-147 (Springer, 2008). https://doi.org/10.1007/978-1-84628-814-2_15

5 M. Hodkiewicz and N. Montgomery, Data fitness for purpose: assessing the quality of industrial data for use in mathematical models, in 8th IMA International Conference on Modelling in Industrial Maintenance and Reliability (MIMAR), (2014). https://www.researchgate.net/publication/295010797_Data_fitness_for_purpose_assessing_the_quality_of_industrial_data_for_use_in_mathematical_models

6 M. Aljumaili, Data Quality Assessment: Applied in Maintenance, Ph.D. thesis, (Luleå University of Technology, 2016)

7 M. Madhikermi, S. Kubler, J. Robert, A. Buda and K. Främling, Data quality assessment of maintenance reporting procedures, Expert Syst. Appli., Vol. 63, pp. 145-164, 2016. https://doi.org/10.1016/j.eswa.2016.06.043

8 M. Rantala, Data quality analysis in industrial maintenance: Theory vs. Reality, Master's thesis, (Lappeenranta University of Technology, 2016).

9 T. Pinna, J. Izquierdo, M. T. Porfiri, J. Dies, Fusion component failure rate database (FCFR-DB), Fusion Eng. Des, Vol. 81, pp. 1391 – 1395, 2006. https://doi.org/10.1016/j.fusengdes.2005.05.011

10 International Organization for Standardization, Nuclear power plants - Reliability data exchange - General guidelines, International standard ISO 6527, 1982.

11 OREDA participants, OREDA Offshore and onshore reliability data handbook, Volume 1, 6th edition, 2015.