Data Quality Assessment on Lecturer Primary Data: A Case Study on Higher Education Database at Ministry of Education and Culture Republic of Indonesia

Rizky T. Prasetyo¹, Yova Ruldeviyani², Editha D. Purnamasari³, Apriliana F. Wibowo⁴

Magister of Information Technology, Faculty of Computer Science, Universitas Indonesia

Department of Information Technology, Faculty of Computer Science, Universitas Indonesia, Jl. Salemba Raya No.4, DKI Jakarta, Indonesia, 10430

Email: ¹rizky.tito91@ui.ac.id, ²yova@cs.ui.ac.id, ³editha.dewi@ui.ac.id, ⁴apriliana.fajri@ui.ac.id

Abstract. The quality of an organization's data is an important component that can influence various policies taken in technical and strategic terms. This study aims to analyze the quality of basic data for lecturers on resource information systems whose data is integrated with the Higher Education Database (PDDikti). Measurement of data quality based on the rules that have been set in the main unit of higher education data management which is then determined in this study includes 4 dimensions of data quality, namely completeness, validity, accuracy, and currency. The application of quantitative methods by processing data and then analyzing the quality of the data using qualitative methods based on the dimensions that have been determined for the basic data quality assessment for lecturer in every tertiary institution in Indonesia. The results showed that the quality of the data had been resolved by 91.94% so that it can be used as a basis for decision making according to predetermined standards. The results of this data quality research can be used as supporting information in policy making in terms of developing and improving data quality in the scope of higher education, especially those related to the quality of basic data for lecturers.

1. Introduction
Data is one of the important components as a driving force for an organization's business that can have a significant impact on achieving the organization's strategic objectives [1]. Knowledge of the quality conditions of data can be an important factor in producing valuable data for the organization. This can be in line with the organization's strategy to determine data quality improvement policies and in terms of maintaining data quality so that it can provide benefits to the organization [2]. Some of the positive impacts of data quality management include increasing the value of organizational data, reducing the risk and cost of poor data quality, increasing organizational efficiency and productivity and increasing the reputation of the organization [1]. To have a positive impact it is necessary to identify the prevailing business processes in an organization that can have an impact on data quality and then be linked to find the root causes of the problem [3].
Various studies on data quality measurement have been carried out in various fields such as in the health sector, education sector to information technology sector. D. Effendi conducted research by measuring the quality of data and information at UNIKOM using the CALDEA and EVAMECAL methodologies [4]. Wijayanti, et al. Conducted a study on the assessment of data quality in higher education using a case study from a tertiary institution in Indonesia [2]. In addition, Herawani also conducted research on data quality improvement strategies covering student and learning activities in the Higher Education Database at the Ministry of Research, Technology and Higher Education [5]. From various studies on data quality, it can be seen that it is important to measure the quality of data from an organization.

Referring to the research on data quality in the Higher Education Database conducted by Herawani which is still in the scope of student and learning data, while in the Higher Education Database there are still several scopes of data that are important to know the quality of the data. In this case, related to the main data of teaching staff who are in the scope of the main unit of the ministry of education and culture organization is the main thing that will be measured the quality of the data. Measuring the quality of data in the context of teaching staff data is quite important to be able to find out how good the quality of the data you have so that it can support policy making by leaders in the main unit to the university level. In addition, there are still reports, especially from data owners, namely lecturers regarding the correctness of data that are not in accordance with the actual conditions, which also underlies the need to measure the basic data quality of teaching staff.

This study aims to analyze the quality of lecturer data, especially the primary data on lecturer registration contained in the Higher Education Database (PDDikti) which is reviewed from several strategic and technical rules such as regulations regarding the governance of the Higher Education Database or regulations regarding National Higher Education Standards as well as other regulations relating to the management of higher education data management in Indonesia. The methods used include quantitative and qualitative in measuring data quality. Quantitative method is used as one of the things in obtaining measurement results from data that has been obtained through SQL Query, while qualitative method is focused on analyzing the results of measuring data quality. The application of stages in the concept of Total Data Quality Management (TDQM) is based on the rules for the dimensions of data quality that are processed based on the provisions contained in PDDikti.

2. Literature Review

2.1. Data Quality

Data quality is the level of meeting expectations or expectations of data users on the data generated [6]. These expectations are defined into related terms with data characteristics that ensure data conformity for a specific purpose. Data quality requirements can be outlined in general and in more detail [6]. Requirements load dimensions like accuracy, completeness, validity and etc. Then, from these requirements will be compared with existing data for measured levels suitability, the higher the suitability level it will be the higher the data quality level.

2.2. Data Quality Dimension

The Data Quality dimension is a characteristic used as a reference for measuring the quality of the data. The data experts have provided a set of quality related dimensions data, according to DAMA DMBOK there are 11 (eleven) dimensions related to the data. Although nothing was agreed upon to set the dimensions of data quality, this formulation contains ideas in general [1]. To that end, every dimension is applied to an agency or dependent company at their respective policies.

The following are some data quality dimensions according to DAMA – DMBOK:

- Accuracy, refers to the level of data that is correctly represented the “real life” entity being modeled.
Completeness, refers to an expectation of completeness indicates that a particular attribute always has a value defined in the data set. Complete is when all attribute values that should have been collected have collected [7].

- Validity, refers to whether data instances are stored, exchanged, or presented in a consistent format with the value domain, and consistent with the attribute value other similar.
- Currency, refers to the extent of current information with the world he is modeling.
- Timeliness, refers to the estimated time to accessibility and availability of information.

2.3. Data Quality Management

Data quality management is one of the activities from data governance to manage data quality. These activities include planning, implementation, and control for ensure the data is suitable for consumption and can meet the data needs of consumers. In general, it is applied using a methodology or technique [1]. Several methodologies have been developed to try describe procedures so that data quality can continue improved.

The most popular methodologies in data quality management are Total Information Quality Management (TIQM) and Total Data Quality Management (TDQM) [8]. Both TIQM and TDQM have a similar goal of providing a methodology for improving the data quality continuously. While TIQM is heavily influenced by practical experience, TDQM is the result of several years of research and can propose to identify and document information production processes and information product characteristics [9]. The methodology is applied according to the organization's data needs.

2.4. Data Quality Measurement

Measurement or assessment of data quality is one of the stages of data quality management. In the TDQM methodology, this measurement or assessment stage enters the second stage. The phases owned by the TDQM methodology are:

- Definition Phase, during this phase, the characteristics of the information product (IP) are captured, such as information requirements, core information objects and components, and their relationship [10]. In addition, the importance of the data quality dimension in the perceptions of IP suppliers, producers, consumers, and managers was identified through a survey that captures the first assessment of the quality of the underlying IP [10]. Furthermore, the information generation system is documented through the so-called "information generation analysis matrix" as the basis for further analysis and improvement [10].

- Measurement Phase, this phase is the beginning of developing data quality metrics. Metrics do not have to directly relate to data, but also with production or access processes, e.g. which updates how much data or how many unauthorized access occurred [10]. The developed metrics are implemented in a system and applied to the data to measure the quality of IP data periodically.

- Analysis phase, based on the measurement results, the root causes of identified data quality problems are analysed during this phase [10]. Metrics are also subject to further analysis, because sometimes there are metric results that need to be adjusted, extended, or improved [10].

- Repair or Upgrading Phase, in this phase the identified causes of quality problems need to be eliminated during this phase. Therefore, it is necessary to identify the necessary improvements, such as adjusting information and workflows to the infrastructure or modifying IP characteristics according to business needs [10]. The improvement phase activities are supported by the "information generation analysis matrix", which was originally created during the definition phase [10].

3. Methodology

The research was conducted with a case study of teaching staff data in higher education databases using quantitative data calculation methods and then analyzed using qualitative methods to obtain measurement results from the quality of teaching staff data in the higher education database. Assessment of the quality of the data uses direct SQL queries (Structured Query Language). The data used to measure
the quality of the data is taken with limitations in the form of data for lecturers who are recorded as active, leave, study assignments, study permits and assignments in other agencies until May 2020 in the resource information system database that is integrated with the higher education database. The steps taken in conducting this research include the first three stages contained in the TDQM method. Furthermore, several stages used in this study were compiled and explained as follows:

Stage 1: Identify the required data based on the existing regulations.
Stage 2: Determine the dimensions of data quality that are used to measure data quality. Determination of the dimensions used is based on the Regulation of the Minister of Research, Technology and Higher Education Number 61 of 2016 concerning Higher Education Database [11].
Stage 3: Define criteria according to the rules applied to the assessed data.
Stage 4: Collect data based on the need for a data quality assessment from the database management system (DBMS). At this stage, PDDikti data are prepared related to lecturer data.
Stage 5: Measuring or assessing the quality of the data based on predetermined criteria. The result of measuring the quality of this data is in the form of a metric of the percentage of data conformity with the rules in the main data for lecturers.
Stage 6: Analyze the results of the data quality assessment based on predetermined dimensions. The analysis was carried out by means of observation and interviews with several officials related to the higher education database.
Stage 7: Provide conclusions and recommendations on the results of data quality assessments based on the results of data processing described in the previous stage.

4. Results and Discussion
Measurement of the quality of teaching staff data is carried out using the TDQM method. The data for lecturers is measured based on the Regulation of the Minister of Research, Technology and Higher Education Number 61 of 2016 Article 12 paragraph 2, this regulation is related to the dimensions applied to higher education data consisting of completeness, truth, accuracy, and up-to-date data [10]. The measurement results display the percentage of data conformity from each dimension with existing business rules, as illustrated in Table 1.

**Table 1. Results of Data Quality Measurement**

| Dimension   | The number of business rules | Business rules are met | Business rules that cannot be measured |
|-------------|------------------------------|------------------------|----------------------------------------|
|             | Total | Percentage | Percentage | Percentage |
| Completeness| 86    | 100%       | 93.33%     | 6.67%     |
| Validity    | 13    | 100%       | 88.91%     | 11.09%    |
| Accuracy    | 18    | 100%       | 95.45%     | 4.55%     |
| Currency    | 3     | 100%       | 90.07%     | 9.93%     |

4.1. Measurement for Completeness Dimension
In the completeness dimension, anomalies are traced to the data, by ensuring that the expectation of the necessity of data values for certain attributes can be met. In this study, the attributes that are expected in the data of lecturers contained in the higher education database, as many as 86 attributes are required to be filled in or cannot be null. If the attribute contains zero, ‘-‘ or contains only spaces, the attribute is grouped into attributes that do not meet the criteria for a business rule. The rules for the completeness dimension are:

C1. A Resource entity that includes 23 attributes that cannot be null.
C2. A Certification Entity that includes 15 attributes that cannot be null.
C3. Power entity that includes 15 attributes that cannot be null.
C4. Functional Entity that includes 12 attributes that cannot be null.
C5. A formal education entity that includes 21 attributes that cannot be null.

Figure 1 displays the measurement results for this completeness dimension that has been met at 93.33%. The discrepancy between data and business rules is 6.67%.

![Figure 1. Graph of Measurement Results for Completeness Dimension](image1)

4.2. Measurement for the Validity Dimension

In the validity dimension, anomalies are traced to the data, by ensuring that there are data anomalies that are inconsistent with the domain values that have been made appropriate in the higher education database. In this study, measuring the validity dimension can be done by ensuring that the data already in the database management system (DBMS) is in accordance with applicable business rules. The business rules in the validity dimension are:

V1. Lecturers of diploma one and diploma two programs must have the minimum academic qualifications of a master's degree or applied master's degree.
V2. Lecturers of diploma three and diploma four programs must have academic qualifications at least as a graduate of a master's degree or an applied master's degree.
V3. Lecturers of undergraduate programs must have academic qualifications of at least a master's degree or applied master's degree.
V4. Lecturers of professional programs must have academic qualifications of at least a master's degree or applied master's degree relevant to the Study Program.
V5. Lecturers of master's programs and applied master's programs must have academic qualifications of a doctoral or applied doctoral graduate.

![Figure 2. Graph of Measurement Results for the Validity Dimension](image2)
V6. Lecturers of specialist and subspecialty programs must qualify as subspecialty graduates, doctoral graduates, or applied doctoral graduates relevant to the Study Program.

V7. Lecturers of doctoral programs and applied doctoral programs must have academic qualifications of doctoral or applied doctoral graduates relevant to the study program, and can use professionally certified lecturers that are relevant to the study program and are qualified.

V8. The provision of a registration number adheres to the principle of 1 (one) lecturer only having 1 (one) registration number.

V9. Lecturers who have a NIDN are given a starting number ranging from 00 - 87.

V10. Lecturers based on a work agreement who have a NIDK are given an initial number 88.

V11. NUP for Tutors is given a starting number 97.

V12. The NUP for Instructors is assigned a starting number 98.

V13. NUP for Non-Permanent Lecturers is given a starting number 99.

Figure 2 displays the measurement results for the validity dimension, which has been fulfilled by 88.91% on the measured data for the 13 predetermined business rules, while there are 11.09% for data that are not validated. The following is a percentage graph for data invalidity in the measures that have been carried out.

4.3. Measurement for Accuracy Dimension

On the accuracy dimension anomalies are traced to the data, by ensuring that the data has a consistent, correct, and unambiguous value. In this research, the measurement of accuracy dimension can be done by ensuring that the data already in the database is in accordance with the current rules in the higher education database. The rules for accuracy dimension are:

A1. Year of Graduation is less than 2020
A2. Year of Entry is over 1900
A3. Year of Graduation is more than 1900
A4. Entry year is less than 2020
A5. Graduation Date is less than or equal to 26 May 2020
A6. TMT SK rank is less than or equal to 2020
A7. The date of SK rank is less than or equal to 2020
A8. The working period of the month group ranges from 0 to 12
A9. TMT SK Functional Position is less than or equal to May 26, 2020
A10. ID of functional position between the numbers 40 to 51
A11. The year of certification is less than or equal to 2020
A12. The date of the Assignment Letter is less than or equal to the date of May 26, 2020
A13. TMT Assignment Letter less than or equal to May 26, 2020
A14. The date of the CPNS SK is greater than the date of birth
A15. TMT SK Appointment is greater than the date of birth
A16. Year of Entry is less than the year of graduation
A17. TMT SK Rank is greater than the date of SK Rank
A18. TMT Assignment Letter is greater than the date of the Assignment Letter

Figure 3 displays the measurement results for this accuracy dimension, which is 95.45% of the data measured for the 18 specified business rules, while there are 4.55% for inaccurate data. The following is a percentage graph for data accuracy on the system.

4.4. Measurement for Currency Dimension

In the currency dimension, anomalies are traced to the data, by ensuring that the level of updating of the data matches the data object being modeled. In this study, dimensional measurements can be made by ensuring that the recording of time when saving data changes is adjusted to the applicable data format. The current data format contained in several attributes for teaching staff data [11] [12], as follows:

CR1. PNS NIP with a compiler code consisting of 18 (eighteen) digit numbers.
CR2. NIK with a compiler code consisting of 16 digits or numbers.
CR3. NPWP has a serial code with 15 digits or numbers.

Figure 4 displays the measurement results for this currency dimension as a whole that has been met by 90.07%. Meanwhile, the discrepancy was found to be 9.93%.

4.5. Analysis of Causes of Data Quality Problems
Based on the problems identified in each data quality dimension, several causes of data problems were found. The cause of the problem is traced by conducting observations and interviews with information technology staff in charge of managing the integrated resource information system. Problems that occur in data quality are because:

- In the completeness dimension for attributes which cannot be null, there are still 18 attributes that are empty or '-' or spaces. This is because the data entered through the current system by the end-user has not been maximally implemented, besides validation that is still being carried out by users in the main unit can also cause errors in data verification. Another thing that causes incompleteness of data is still incompleteness of the data on the migration results.
• In the validity dimension, data is found that are validated by the current business rules. This is because the data has not used the latest validation rules. For example, validation for non-permanent lecturers’ NUPs still has a bug in the system which causes the number given by the system to users not validated in accordance with the applicable rules.

• In the Accuracy dimension, data is found that are inaccurate and not in accordance with the rules previously stipulated in Regulation of the Minister of Research, Technology and Higher Education Number 61 of 2016 Article 12 paragraph 2 [13]. This is due to the lack of awareness of users to update data, besides the accuracy of checking by validators also affects the level of data accuracy so that it cannot be known that the data of lecturers has been updated massively.

• In the Currency dimension, data is found that have not been latest or updated on the attributes of NIDN, NIK, NIP and NPWP. This is due to several factors, such as errors when the system gave the lecturer registration number to lecturers. In addition, regulatory updates from the bureaucracy can also affect data changes that must be made, especially in database structures.

5. Conclusion
Based on the results and data analysis on the research that has been carried out on the main data of lecturer registration at PDDikti, the whole data has been fulfilled by 91.94% with compliance for the Completeness dimension of 93.33%, Validity 88.91%, Accuracy 95.45% and Currency 90.07%. The data anomalies found in this study are caused by several things such as awareness of updating data for each user, the ability of validator resources in the main unit, and some adjustments in the system related to bugs found and policy changes that affect data changes and database structures from higher education database.

Adjustments to the database structure need to be made in relation to adjustments to the bureaucratic structure of higher education which is back under the Ministry of Education and Culture. In addition, it is necessary to socialize the importance of maintaining data quality to each user from the university level to the validator user in the main unit according to the needs of PDDikti.

This study has limited data in terms of measuring data quality. The data used includes data on teaching staff which includes basic data on lecturer registration at PDDikti. The data taken includes data for teaching staff who are active, study assignments, study permits, leave and other agency assignments up to the deadline of May 2020 which are recorded in the resource information system integrated with PDDikti. This is caused by licensing issues from the organization regarding data security.

This research only reached the third stage of the TDQM methodology, namely the analysis of the causes of data quality problems in a case study of primary data on lecturer registration at PDDikti. For this reason, it is necessary to have research into the next stage of TDQM in order to be able to make strategies in improving the quality of basic data on teaching staff owned by PDDikti. In connection with the rules that have been made regarding each dimension of data quality, analysis of several documents contained in PDDikti and PP Number 61 of 2016 Article 5 paragraph 1. Furthermore, this research can be reviewed and re-analyzed by paying attention to several regulations issued due to bureaucratic restructuring in higher education. In addition, this research can be used as a reference for other people who want to evaluate the quality of higher education data from universities in Indonesia that follow the PDDikti requirements in accordance with the applicable rules in PDDikti and other related regulations.

References
[1] Dama International 2017 DAMA - DMBOK Data Management Body of Knowledge, Second Edition (New Jersey: Technics Publications)

[2] Wijayanti W, Hidayanto A N, Wilantika N, Adawati I R and Yudhoatmojo S B 2018 Data quality assessment on higher education: a case study of institute of statistics Int. Seminar on Research of Information Technol. and Intelligent Systems (ISRITI) pp 231–6

[3] Loshin D 2011 The Practitioner’s Guide to Data Quality Improvement (USA: Elsevier)
[4] Effendi D 2012 Measurement and improvement of data and information quality in higher education using CALDEA and EVAMECAL Seminar Nasional Rekayasa Sains Dan Teknologi (RESATEK) 2 2087–526

[5] Herawani P 2016 Data Quality Improvement Strategy: A Case Study of Higher Education Database at The Ministry of Research, Technol., and Higher Education (Jakarta: Universitas Indonesia)

[6] Coleman and Sebastian L S 2013 Measuring Data Quality for Ongoing Improvement A Data Quality Assessment Framework (USA: Elsevier)

[7] Liu L and Chi L 2002 Evolutionary data quality: a theory-specific view Proc. of the 7th Int. Conf. on Information Quality pp 292–304

[8] Huang K, Lee Y W and Wang R Y 1999 Quality information and knowledge (USA: Prentice Hall)

[9] Fürber C 2015 Data Quality Management with Semantic Technologies (Munchen: Springer)

[10] Wang R Y 1998 A product perspective on total data quality management Communications of the ACM 41 pp 58–65

[11] Kemenristekdikti 2016 Appendix Number 2 of 2016 concerning Amendments to the Regulation of the Minister of Research, Technology and Higher Education Number 26 of 2015 concerning Registration of Lecturers in Higher Education [in Indonesian] (Jakarta: Kemenristekdikti)

[12] BKN 2007 Regulation of the Head of BKN Number 22 of 2007 concerning Civil Servant Identity Number [in Indonesian] (Jakarta: BKN)

[13] Kemenristekdikti 2016 Regulation of the Minister of Research, Technology and Higher Education Number 61 of 2016 on Higher Education Database [in Indonesian] (Jakarta: Kemenristekdikti)