Data Quality and Data Gaps in Educational Data Base (U-DISE)

Sheeraz Ahmad Peerzada[1] , Dr.Jitendra Seethalani[2]

[1-2]Department of Computer Science,
Sri Satya Sai University of Technology and Medical Sciences,
Sehore, MP, (India)

ABSTRACT

All organizations either government or private confront on data quality. Data quality is the key of success for any organization. The Department of School Education stores all information in Universal data base called Unified District Information System for Education (U-DISE).Since U-DISE is a heterogeneous data base, stores information related to students, teachers, schools and facilities etc. Data Gaps must analyze and removed from this Universal Data Base (U-DISE) to achieve high quality. High quality data can be used in operations, decision making and planning. The main focus of this paper is to understand data quality and data gaps of Educational data base "Unified district information system for Education" (U-DISE).

Keywords-Data quality, Parameters, Educational Data Base (U-DISE), Data Gaps.

I.INTRODUCTION

Data quality means the degree of excellence exhibited by the data in relation to the actual scenario. Production of high quality statistics depends on the assessment of data quality, without good approaches for data quality assessment statistical institutions are working in the dark.

Data quality is the ability of a given data set to serve an intended purpose. To put it another way, if we have data quality, the data is capable of delivering the insight we hope to get out of it. Conversely, if we do not have data quality, there is a problem in our data that will prevent us from using the data to do what we hope to achieve with it.

The Department of Education is collecting information on yearly basis in India, and last reference date is 30th of September every year. The said data is stored in a nationalized Data Base called Unified District Information System for Education (U-DISE).

Educational Data Mining is yielding good results on this U-DISE data, but the data must be qualitative not quantitative data.

U-DISE software is operational all over India. The Central Government has launched U-DISE for collection statistical data, every school must fill information on U-DISE data capture format yearly. Data quality is essential for this data base the U-DISE data is used for decision making and annual works plan formation (AWP & B).The data is also used for various activities of SSA & RMSA .

II.LITERATURE REVIEW ON DATA QUALITY

In the 1950s, researchers began to study quality issues, especially for the quality of products, and a series of definitions, for example, quality is “the degree to which a set of inherent characteristics fulfill the requirements” (General Administration of Quality Supervision, 2008); “fitness for use” (Wang & Strong, 1996); “conformance to requirements” (Crosby, 1988) were published. Later, with the rapid development of information technology, research turned to the study of the data quality. Research on data quality started abroad in the 1990s, and many scholars proposed different definitions of data quality and division methods of quality dimensions. The Total Data Quality Management group of MIT University led by Professor Richard Y. Wang has done in-depth research in the data quality area. They defined “data quality” as “fitness for use” (Wang & Strong, 1996) and proposed that data quality judgment depends on data consumers. At the same time, they defined a “data quality dimension” as a set of data quality attributes that represent a single aspect or construct of data quality. They used a two-stage survey to identify four categories containing fifteen data quality dimensions [1].

2.1 Global Data Quality Research

Prevent problems before they occur Retrospective cleaning of data has always existed; usually after the data quality problem has snowballed into an issue. Often at this stage, monetary or reputational damage may also have occurred. Organisations need to look at their data quality processes and determine if the problem can be turned on its head, and tackled at the front end. Email and mobile data is a common example. I often get to profile customer data and these two data entities have collected a lot of garbage data, such as test@test.com for emails and 000000000 for mobile numbers. Today, we have technologies that check if the email or mobile numbers are structurally valid and you can go a step further to check if these actually exist, thus ensuring the data is captured correctly the first time. Using the healthcare analogy, think of it as a vaccination for your quality ailments, rather than popping pills repeatedly for a chronic condition, this can be expensive in the long-term. When improving your level of data sophistication is your goal, then people, processes and technology are the building blocks you need to get there. Improving data quality isn’t a one-off task with an end-point. Instead, it needs long-term investment and commitment, to ensure your data remains both accurate and appropriate for your business needs.[2]

Sheeraz Ahmad Peerzada,Zubair Ahmad and Dr. Jitendra Seethlani published a paper on IJARSE “Role of Data Mining on Educational Data Bases” were they suggested for each indicator different Data Mining algorithms may be applied and the algorithm with optimal results may be used for improving overall performance of students and look in to week areas of students as well as institutions.[3].

III. PARAMETERS OF DATA QUALITY

Data quality is following parameters accuracy, completeness, consistency and validity.

3.1 Accuracy-Accuracy refers to the closeness of a measured value to the standard or true value.

3.2 Completeness-Data is fully filled and no information should be blank.

3.3 Consistency-The degree to which values are present in the attributes that require them. Not changed and match with some parameters.

3.4 Validity-Same information if collected by any one.

IV. WHERE FROM DATA GAPS OCCURS IN U-DISE DATA.

All though U-DISE data maintains 75% accuracy but still lagging 100% accuracy and 5% sample checking is done by Statistics and Evaluation Department.

Common errors-

4.1. At the time of DCF filling- number of teachers, classroom, infrastructure facility, School building and enrollment (SC/ST).

4.2. At the time of Data entry-data entry: 19-91,1-10, extra digit, wrong school, school missed.

4.3. Data format has not properly checked at all levels.

4.4. No supervisory machinery available and checking by others.

4.5. Blank information missed critical information and no response.

4.1 However there are several other gaps described below.

4.1.1 Some fields in DCF are ambiguous, misinterpretation by respondents.

4.1.2 All states and districts face time and resource challenges.

4.1.3 The mainly focus is usually on timely completion of the data collection for use in annual planning process.

4.1.4 Most administrative (non-technical) staff are dependent on the MIS staff for accessing the data.

4.1.5 The present infrastructure in schools is not sufficient were schools can submit the DCF electronically.

4.1.6. Some data is collected too frequently and some data that is essential not collected at all [4].

V.QUALITY ASSESSMENT FOR U-DISE DATA

An appropriate quality assessment method for U-DISE data is necessary to draw valid conclusions. In this paper, we propose an effective data quality assessment process with a dynamic feedback mechanism based on U-DISSE data’s own characteristics, shown in **Figure 1**.

Determining the goals of data collection is the first step of the whole assessment process. U-DISE data users rationally choose the data to be used according to their strategic objectives or academic requirements, such as operations, decision making and planning. The data sources, types, volume, quality requirements, assessment criteria, and specifications as well as the expected goals need to be determined in advance.

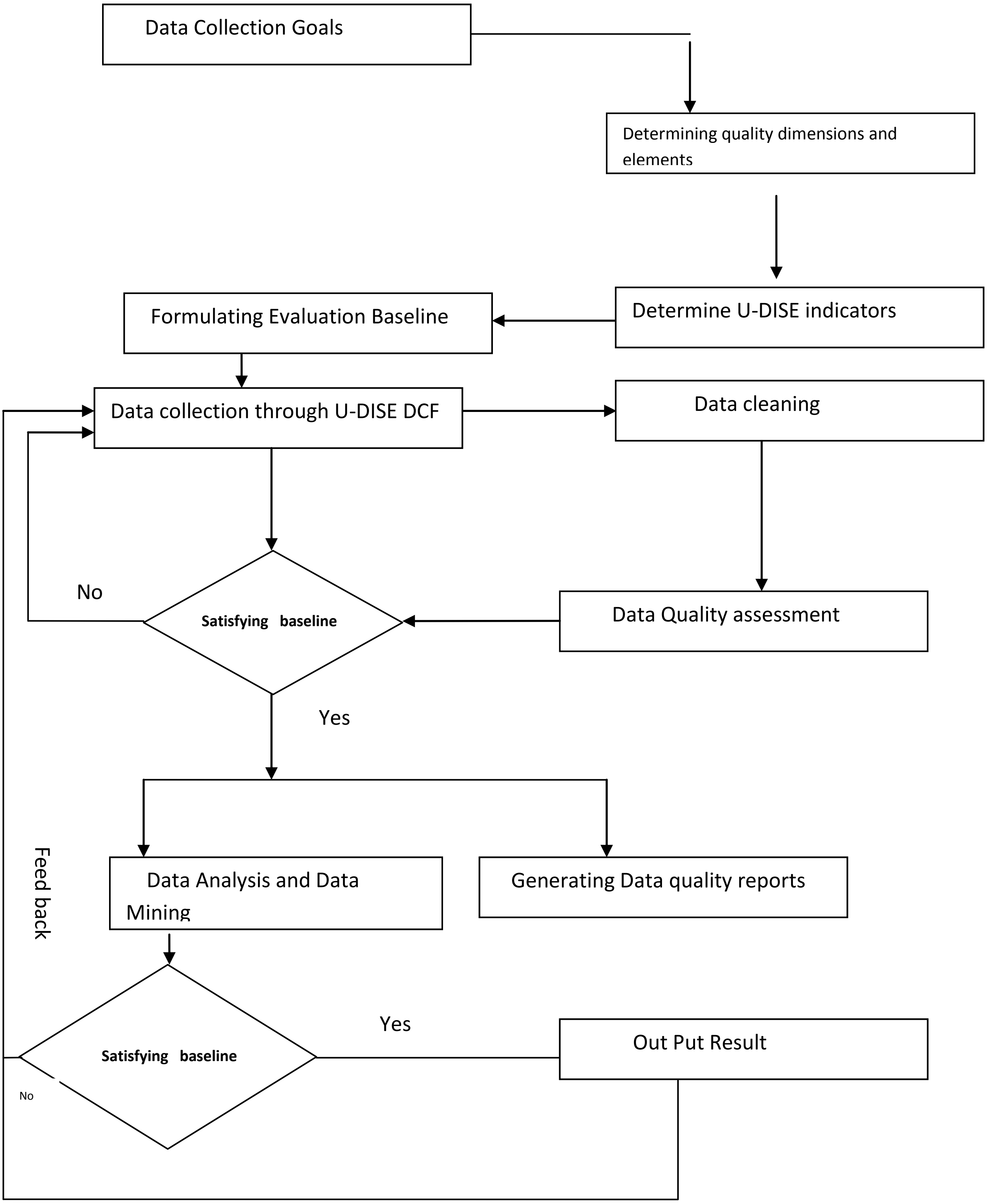
Data Collection Goals
Determining quality dimensions and elements
Determine U-DISE indicators
Formulating Evaluation Baseline
Data collection through U-DISE DCF
Data cleaning
No
Satisfying baseline
Data Quality assessment
Yes
Data Analysis and Data Mining
Generating Data quality reports
Feed back
Yes
Satisfying baseline
Out Put Result
No

Fig 1. Quality Assessment Process for U-DISE Data.

IV.CONCLUSION

The increased use of Technology in Education leads arrival of big data, the data from students, teachers as well as from educational institutions. To ensure data quality, data gaps must be analyze and removed. Poor data quality will lead to low data utilization efficiency and even bring serious decision making mistakes we analyzed the challenges faced by the Department of School Education during collection and implementation of national Data base (U-DISE).We formulated a dynamic data quality assessment process with a feedback mechanism .The further research in this field is to design an algorithm for data quality.

V.ACKNOWLEDGEMENTS

I am thankful to my research guide Dr.Jitendra Seethlani for his suggestions, considerations and his guidelines which he provided me day to day. His responsive advices help me to work better and quality improvement.

REFERENCES

[1] Li Cai,and Yangyong Zhu , “ The Challenges of Data Quality and Data Quality Assessment in the Big Data Era”.

[2] Global Data Quality Research Discussion Paper 2015.

[3] Sheeraz , Zubair Dr. Jitendra Seethlani published a paper on “Role of Data Mining on Educational Data Bases” , IJARSE Journal 2017.

[4] U-DISE data quality assessment by RMSA task group “a review of the U-DISE data on selected indicators”