Improving Data Quality in Crowdsourced Data for Indonesian Election Monitor: A Case Study in KawalPilpres

F Gunawan and Y Ruldeviyani
Faculty of Computer Science, University of Indonesia, Jakarta, Indonesia
Email: fandi.gunawan@ui.ac.id, yova@cs.ui.ac.id

Abstract. ICT has enabled democratic process to be more transparent and enabled citizens’ participation in the election process. However, public trust is a mandatory requirement for a good democratic transition. Participation in monitoring election process could be implemented as a crowdsourcing effort to improve public trust in the election result which in this article based on a case study of KawalPilpres to monitor 2019 Indonesian presidential election. Trust factor is a key success for monitoring effort. Therefore, data quality becomes necessity. Data quality is assessed using Loshin’s maturity assessment and analysed using Loshin’s improvement strategies. Based on our assessments, there are three top categories for improvements namely governance, expectations, and policies of data quality management.

1. Introduction
Democracy entails citizens participating in the democratic process and to ensure the quality of the democratic process followed. The quality of democratic process includes freedom, equality, and responsiveness [1]. The characteristics of the quality require citizens participation beyond voting in the election [1]. Crowdsourcing is activities that include various non-homogenous, distributed, and participative activities that bring benefit and satisfaction among people who are in the activities [2]. Crowdsourcing, particularly within election has been a central issue of election process transparency and participation. Several examples of crowdsourcing processes such as election in Russia [3], Uganda [4], Mexico [5], and Indonesia [6] have been benefited by the application of ICT. Crowdsourced data has a very high possibility to improve timeliness, quality, and breadth of data [7]. Despite those benefits, there are also issues related to cost, accuracy, ethical challenge, political problems, and legality [7].

Previous researches in crowdsourcing for election monitoring include participation case study [3,4,6,8]. [8] shows the correlation between crowdsourcing and voter turnout. [4] incurs the reasons of user participation while [3] displays the effectiveness in uncovering fraud. [6] discussed citizen engagement perspective in crowdsourcing. While previous researches have shown the positive impacts of crowdsourcing in election process, there is quality aspect that should be explored. [9] suggested the use of auditing system to improve data quality while [10] used special framework. A more technical approach such as probability matrix [11], noise correction [12], and statistical distribution [13]. The organizational or program aspect of data quality improvement are issues that has not been addressed in-depth in previous researches. This research focused on improving data quality in program and organization level.

KawalPilpres is an ad-hoc crowdsourcing program to monitor 2019 presidential election in Indonesia. It was initiated by various religious and democratic watch communities. It utilized highly
secured ICT as a backbone for election monitoring by involving volunteers to capture D day election result, and to perform verification of those result [14]. It had challenges in gathering volunteers, gathering data across Indonesia, limited time in verification process, and data quality issues. There were 508,491 election report from 9,550 volunteers with 336,532 unique voting stations. The result accounted for 41.53% of total national and foreign voting stations. However, there were 164 indications of invalid reports. Despite the small numbers of invalid reports, invalid result might get published in internet to distrust the system as a whole. Another issue was the myriad number of reports requires average verification time of 8.23 days. Data timeliness was also crucial part due to real-time result that should be published into monitoring dashboard and tabulation process. Finally, knowledge gaps among volunteers were also a challenging issue which correlates to data quality. Therefore, this article focuses on how to improve data quality on crowdsourced data for ad-hoc voluntary program. To answer the question, case study research with qualitative approach using data quality framework was selected to improve KawalPilpres.

2. Data Quality in Crowdsourcing Context

2.1. Data Quality

Data quality has been a very important concern in every organization that deals with data. While there is no single definition of data quality exist, it is keen to see that data quality includes correctness of data and fitness for use [15–18]. Data quality may improve business performance [16], help serving customer better [17], and increase customer engagement [6]. In the context of crowdsourced data, data quality takes very important part due to the nature of the process of how the data is being collected and processed [19].

Figure 1. (a) Comparison of election tabulation and KawalPilpres (b) reporting to verification time.

KawalPilpres main data are collected by volunteers by sending photo and data from C1 in voting station and validated by moderators that were selected from volunteers themselves. The result from volunteers were then moderated before it was published and aggregated. Figure 1(a) shows comparison of KPU (National Election Committee) and KawalPilpres process. KPU performs hierarchical manual aggregation from voting station level to national level. Documents C1, D1, DA1, DB1, DC1, and DD1 indicate level in election process from voting station to national level. KPU made the C1 data available for general public. However, its system for C1 calculation had faced scrutiny from competing parties and their supporters.

Table 1. Common problems in KawalPilpres.

| Table Name  | Table description                      | Dimension | Criteria            | Example                                |
|-------------|----------------------------------------|-----------|---------------------|----------------------------------------|
| pelaporanv2 | Election report by volunteers          | Accuracy  | Readable C1         | Submission of invalid C1               |
| relawan     | Volunteers’ information                | Completeness | Volunteer information are complete | Missing e-mail field or user id       |
| verifikasi  | Verification information               | Timeliness | Defined maximum time for verification | There is no defined maximum time for verification |

There were issues during data collection by volunteers and processing that can be categorized into inconsistency, incompleteness, and inaccuracy (table 1). KawalPilpres mitigated some of the issues by
upgrading the system, normalizing the data, and blocking some malicious submissions. The main challenge during data acquisition is inaccuracy due to no standardization of how picture was taken, e.g. clipped C1, non-C1 picture, and low-quality picture. The challenges during processing are inadequate moderators to monitor the quality of data and longer time to process from C1 submission to publication in dashboard. Figure 1(b) indicates the time fluctuation needed from report to verification process.

2.2. Data Quality Assessment and Improvement
Data quality assessment is a process to get the insight and measurement of current data practices. Assessment might be conducted subjectively or objectively by utilizing quantitative or qualitative approach [20]. There are generalized process for data quality assessment which includes defined metrics or combination of metrics, and subjective measurements.

Maturity model was designed to give an insight of certain process that should be controlled, measured, and improved [21]. Maturity model becomes one integrated part for process improvement framework [22]. Despite of its initial purpose to be used in software development process, maturity model is also used in other fields including data quality. There are several data quality maturity model e.g. [15], [23], [24], [25], and [26]. Despite differences on how to do improvements, most maturity models provide improvement by following defined steps. Loshin’s Data Quality framework [15] is chosen due to its detailed steps using maturity model [27] in quality improvement and metric to measurement. Loshin provides step by step guideline for quality improvements in organization.

Data quality improvement is a strategy and steps required to improve the data quality by employing proven methodology and considering the organizational context [20]. Loshin [15] suggests that data quality cycle includes (a) identification and measurement, (b) analysis, (c) design quality improvement, (d) implementation improvement, and (e) monitoring. Loshin’s Data Quality framework assessment follows maturity model in which levels of data quality is in line with its maturity level. There are eight components and five levels namely Level 1 Initial, Level 2 Repeatable, Level 3 Defined, Level 4 Managed, and Level 5 Optimized. Components consist of (a) expectations, (b) dimensions, (c) policies, (d) procedures, (e) governance, (f) standards, (g) technology, and (h) performance management. Figure 2(a) shows theoretical framework that includes democracy quality [1], crowdsourcing [2–9], and data quality [10–13] that become the basis for voluntarily ad-hoc effort for election monitoring.

3. Methodology
Preparation of the research is gathered through interviews to the stakeholders of KawalPilpres and analyzing the data of KawalPilpres. There are two stakeholders interviewed which handle development and data of KawalPilpres. Two stakeholders have more than 10 years experiences in application development and managing projects in IT. Another source of data is database and system of KawalPilpres. From the steps above, a problem statement is defined to improve data quality in KawalPilpres.

Loshin’s data quality assessment includes maturity level as a baseline of data quality assessment [15]. The maturity levels are indicators of the current data quality state. It is an indication to explain the gaps of the existing condition to the next level. Data quality was assessed using interview with 134 closed ended questions from [15] to two stakeholders in KawalPilpres that oversee the system.

Figure 2. (a) Theoretical framework (b) methodology.
development and data. 134 questions were generated from Loshin’s maturity indicators across eight components. Every question is graded using FPNNA, averaged for every level, and then aggregated to get level of certain components. FPNNA scale consists of F (fully performed), P (partially performed), N (not performed), and NA (not applicable). Conversion value of the scale is as follow F=1, P=0.5, N=0, and NA = not calculated. Analysis and improvement recommendations are suggested through process of validating the result of analysis to the stakeholders. The process analysis includes pareto analysis and root causes analysis. Figure 2(b) shows the methodology used in this research.

4. Result
Assessment results were aggregated and averaged using FPNNA that resulted in heatmap figure 3(a) which indicates areas that require improvements. Heatmap is divided into three colors: red, yellow, and green that corresponds to 0, 50 quartiles, and 1. The result shows maturity level is 2.68 where components of data quality expectations, policies, and governance are among the lowest. The organizational objective is to have at least level 3 Defined.

| Components    | 1     | 2     | 3     | 4     | 5     | Total |
|---------------|-------|-------|-------|-------|-------|-------|
| Expectations  | 0.33  | 0.33  | 0.00  | 0.50  | 0.38  | 1.54  |
| Dimensions    | 0.83  | 0.50  | 0.83  | 0.33  | 0.00  | 2.50  |
| Policies      | 0.67  | 0.67  | 0.33  | 0.00  | 0.00  | 1.67  |
| Procedures and Protocols | 0.13  | 1.00  | 0.50  | 0.83  | 0.83  | 3.29  |
| Governance    | 0.50  | 1.00  | 0.00  | 0.00  | 0.00  | 1.50  |
| Standards     | 0.00  | 1.00  | 1.00  | 1.00  | 0.75  | 3.75  |
| Technology    | 0.00  | 0.67  | 1.00  | 1.00  | 1.00  | 3.67  |
| Performance Management | 0.00  | 1.00  | 0.67  | 0.83  | 1.00  | 3.50  |
| Average level of all components |       |       |       |       |       | 2.68  |

Figure 3. (a) Assessment result and (b) radar chart.

Figure 3(b) indicates the gaps within current data quality state and the organizational expectations. There are four components in assessment result that do not meet the expectations, namely (a) expectations, (b) dimensions, (c) policies, and (d) governance. Priorities of data quality improvements are analyzed using pareto chart where 80% of effect are caused by 20% of causes [15,28]. Priorities are collated by reversing average level value of each components.

Figure 4(a) demonstrates three main components that needs to be addressed namely (a) governance, (b) expectations, and (c) policies. Figure 4(b) shows root cause analysis that was selected to identify the root cause of problems in governance, policies, and expectations to achieve minimum level 3. Root causes analysis is a tool selected to identify the main causes of certain problem [15,16,28].

5. Discussion
Assessment found out that governance, expectations, and policies are components that should be prioritized for improvements. [16] indicates governance is an activity that oversees all aspects of data
management functions. Governance is in place to ensure the business goals and constraints are met by ongoing or planned processes [15]. In an ad-hoc organization such as KawalPilpres, it is necessary to have at least data quality charter, roles and responsibility, and operational structure. Roles and responsibility related to data quality should be embedded into organizational structures despite of its limited operational time. Data governance effort should utilize best practice in data quality, and should consider (a) security and privacy, (b) time, and (c) cost constraints. [15,16] include best practices by entrancing top-down approach (a) creation of steering committee to control data quality, (b) creation of data charter, (c) setting roles and responsibilities, and (d) setting operational structures and incorporate data stewardship.

Policies are set statements that are derived from governance to provides rules to data activities [16]. The challenges in an ad-hoc organization is time constraints, therefore policies must be set with simplicity and practical approach in mind, and be embedded into organizational members. The challenges include (a) variation in volunteers’ capability and knowledge gap, (b) IT capability to enforce policies, and (c) organizational support in enforcing those policies.

Expectations are characters that should align between data quality and business objectives [15]. Data expectations should be defined so that public trust to the system and to the whole elections process is improved. The expectations should include relevant dimensions measurement, evaluation metrics, and process and service for evaluating data expectations. Figure 5(a) specifies data quality improvements by aligning data governance, policies, and expectations into organization and its information system. Data management team structure consists of governance and stewardship function (figure 5(b)). Governance body is embodied within KawalPilpres Steering Committee. Data stewardship function consists of (a) data architect, (b) data developer and quality, and (c) data operation and security.

6. Conclusion
Data quality maturity level for KawalPilpres is 2.68 and there are top three components for improvements from pareto analysis namely governance, policies, and expectations. The improvements recommendations should be able deliver a better data quality in the next election monitor project. The current research has limitation of benchmarking the existing condition and conditions where improvements steps are implemented. The future research should be able to perform a longitudinal research in data quality using crowdsourced data with focus in governance, policies, and expectations.

7. References
[1] Diamond L J and Morlino L 2004 The Quality of Democracy : An Overview J. Demoer. 15 20–31
[2] Estellés-Arolas E and González-Ladrón-De-Guevara F 2012 Towards an integrated crowdsourcing definition J. Inf. Sci. 38 189–200
[3] Bader M 2013 Crowdsourcing election monitoring in the 2011-2012 Russian elections East Eur. Polit. 29 521–35
[4] Hellström J and Karefelt A 2012 Participation through mobile phones: A study of SMS use
during the Ugandan general elections 2011 ACM International Conference Proceeding Series (New York, New York, USA: ACM Press) pp 249–58

[5] Salazar O and Soto J 2011 How to Crowdsourc Election Monitoring in 30 Days: the Mexican Experience Mobile Technologies for Conflict Management (Dordrecht: Springer Netherlands) pp 55–66

[6] Purwanto A, Zuiderwijk A and Janssen M 2018 Citizen engagement in an open election data initiative: A case study of Indonesian’s “Kawal Pemilu” ACM International Conference Proceeding Series (New York, New York, USA: ACM Press) pp 1–10

[7] Garcia-Molina H, Joglekar M, Marcus A, Parameswaran A and Verroios V 2016 Challenges in Data Crowdsourcing IEEE Trans. Knowl. Data Eng. 28 901–11

[8] Bailard C S and Livingston S 2014 Crowdsourcing Accountability in a Nigerian Election J. Inf. Technol. Polit. 11 349–67

[9] Misra A, Gooze A, Watkins K, Asad M and Le Dantec C 2014 Crowdsourcing and its application to transportation data collection and management Transp. Res. Rec.

[10] Meek S, Jackson M J and Leibovici D G 2014 A flexible framework for assessing the quality of crowdsourced data 17th Agdl. Conf. Geogr. Inf. Sci. 3-6 June 2014 3–6

[11] Jung H J and Lease M 2012 Improving quality of crowdsourced labels via probabilistic matrix factorization AAI Workshop - Technical Report

[12] Zhang J, Sheng V S, Li T and Wu X 2018 Improving crowdsourced label quality using noise correction IEEE Trans. Neural Networks Learn. Syst.

[13] Paulheim H and Bizer C 2018 Improving the quality of linked data using statistical distributions Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications

[14] Purbo O W, Deograt G, Satriyanto R, Ferdinand A, Gesang P, Kesuma R and Silaen K E 2019 KawalPilpres2019: a highly secured real count voting escort architecture TELKOMNIKA (Telecommunication Comput. Electron. Control. 17 2834–41

[15] Loshin D 2011 The practitioner’s guide to data quality improvement (Morgan Kaufmann)

[16] DAMA-DMBOK 2009 The DAMA Guide to The Data Management Body of Knowledge

[17] Sedigh S 2013 Handbook of data quality: Research and practice

[18] ISO 2015 ISO 8000-8:2015 Data quality — Part 8: Information and data quality: Concepts and measuring (Geneva, Switzerland)

[19] Shoemaker D, Woody C and Mead N R 2016 Advances in software engineering and software assurance Advances in Computers

[20] Cichy C and Rass S 2019 An overview of data quality frameworks IEEE Access 7 24634–48

[21] Humphrey W S 1988 Characterizing the software process: a maturity framework IEEE Softw. 5 73–9

[22] CMMI Product Team 2010 CMMI® for Development, Version 1.3 (Hanscom AFB, MA)

[23] Baškarada S, Koronis A and Gao J 2006 Towards a capability maturity model for information quality management: A TDQM approach Proc. 2006 Int. Conf. Inf. Qual. ICIQ 2006

[24] Caballero I, Caro A, Calero C and Piattini M 2008 IQM3: Information quality management maturity model J. Univers. Comput. Sci. 14 3658–85

[25] Ryu K S, Park J S and Park J H 2006 A data quality management maturity model ETRI J. 28 191–204

[26] Kurniati A and Surendro K 2010 Designing IQMM as a Maturity Model for Information Quality Management Proc. 2010 InSITE Conf. 277–90

[27] Sabtiana R, Yudhoatmojo S B and Hidayanto A N 2019 Data Quality Management Maturity Model: A Case Study in BPS-Statistics of Kaur Regency, Bengkulu Province, 2017 2018 6th International Conference on Cyber and IT Service Management, CITSIM 2018

[28] Pyzdek T and Keller P 2010 The Six Sigma Handbook (New York, NY: McGraw-Hill)