ODK-X: From A Classic Process to A Smart Data Collection Process

https://doi.org/10.3991/ijim.v15i13.22945

Iman Tikito (✉)
Mohammed V University, Rabat, Morocco
iman.tikito@gmail.com

Nissrine Souissi
Mines-Rabat School, Rabat, Morocco

Abstract—Data collection is one of the first and main phases of the data life cycle. It enables improvements to be made across all phases of the data lifecycle. In this sense, we have proposed a data collection process qualified as Smart. For our smart data collection process, we have adopted the principles of the smart data approach allowing less data to be transmitted to the analysis and storage processes, while maintaining better data quality. In addition, we also used Edge computing since it provides services with faster response and better quality, compared to cloud computing. To experiment this process on mobile data, we propose to extend a mobile data collection software solution and adopt one of the key data collection methods. In this paper, we tested our smart data collection process via the ODK-X software suite and were able to identify the added value of our process compared to the one used by default during collection.

Keywords—Smart Collect, Mobile Data, ODK-X, Data collection, Edge computing

1 Introduction

To experience the smart collection process [1] on mobile data, we come to implement it by extending a software solution for mobile data collection and adopt a data collection method among the five key methods of data collection which are presented with their advantages and disadvantages in article [2]. These five methods are: Surveys, Interviews, Focus Groups, Observations, and Textual or Content Analysis. Thus, our choice was based on the use of a survey to document perceptions, attitudes, and knowledge within a clear and predetermined sample of individuals.

The implementation aims to demonstrate Smart collection process added value [3] compared to a classic data collection process [4], [5]. We are particularly interested here in the collection of mobile data. To minimize cost and workload while being efficient in providing better data quality, we adopted Electronic Data Collection (EDC) to establish the survey as the best approach in our case according to several articles [6], [7], [8].

Depending on the space and quantity of units covered by the survey, gathering and entering the data collected requires appropriate logistics [9], [10], [11]. Now with the use of new technologies like smartphones and tablets in data collection, researchers has no longer need to use paper formats for data collection. Electronic data collection significantly reduces the time between collection and decision-making by eliminating the step of entering the data collected. Thus, the data can be transmitted automatically to databases once collected. Aside from saving time and money, electronic collection of survey data allows real-time monitoring of the collection process progress [12].

There are a variety of tools and digital platforms that facilitate the collection of data from smartphones and tablets; some are more effective than others. The study presented in CartOng [13] raises 26 tools dedicated to electronic data collection. Through the comparison between the described tools, choosing ODK-X as a tool is clearly the best choice for our needs [14], [15].

The Open Data Kit is a community of developers, staff within institutions and organizations developing Open Data Kit for collect, manage and use data in resource-limited settings. ODK is deployed as a collection of GitHub repositories that anyone can use under the Open-Source license. New or old versions are provided through the GitHub portal. Transparent and open problem management is carried out by the Open Data Kit alliance. ODK consists of 68 packages covering different features of ODK and is constantly improving [16].

The survey carried out in this paper concerns an engineering school wishing to know the home establishment (CPGE) of students who obtained good grades in the first year of school while allowing an analysis by gender, by CPGE, etc. This survey concerns only students enrolled in the school. Thus, it is requested to create a form containing the necessary fields to collect the appropriate data.

This article is organized into 4 sections. Besides the introduction, the second section begins with a presentation of ODK-X software suite, then conduct our case study according to the classic collection process thus the proposed Smart collection process. Then, the third section present the results of the two approaches before concluding the article in forth section.

2 Research Method

The Open Data Kit (ODK), as mentioned in several articles like [13], [17] and [16] is one of the best-known software suites in this field and many researchers continue to improve it in order to better meet the various user expectations. It is easy to learn and no extensive training is required to handle it. It was designed to be used by anyone with or without programming skills, thanks to its operation sufficiently adapted to a large population. It consists of a suite of mobile data collection tools and has desktop and mobile applications for data collection and management as well as a server for synchronizing the collected data.

The objectives of ODK as mentioned in [18] are:

- Make the tools modular and customizable so that they can be easily interoperable for each deployment.
- Exploiting interfaces and open standards so that solutions are not "compartmentalised" in monolithic packages of difficult company to understand and maintain.

Following the article [19], the use of ODK is constantly increasing and this is linked to the fact that it is free and open source allowing to appropriate the tool with good programming knowledge, otherwise a simple use of its wide range is also sufficient for some cases. However, the reason for its popularity is also linked to its online support community which is active and constantly improving the various components.

On their official website [20], the community has released two suites ODK and ODK-X that coexist and that one does not replace the other. Each suite contains tools that work together to collect, use, and manage data, but the two suites require different levels of technical skills. In general, ODK tools are easier to use, require less configuration, and are widely adopted. However, for a complex study and with technical skills, ODK-X tools may be better suited.

Our choice fell on ODK-X for the following reasons:

- Flexible tool suite that supports complex workflows through JavaScript customization
- Non-sequential navigation
- Bidirectional synchronization
- Data management on the device

One of the main goals of ODK-X tools is to reduce the complexity that organizations encounter with the software engineering skills required when designing data management applications. ODK-X allows developers and data managers to build data management applications that consist of survey forms as well as JavaScript-based applications as needed. Organizations generally use productivity software such as Excel to create these applications. Thus, the skills required to create a data management application are based on writing a form definition in XLSX then processed by the XLSForm tool, or simple web programming using HTML and JavaScript for personalized presentations. Advanced web programmers can easily implement fully customizable web pages to collect, manage and visualize data on an Android device [20].

ODK-X [19] enables the creation and customization of domain independent mobile applications that meet the needs of an organization within the constraints imposed by Android. ODK-X's protocols and timing structures are designed to be adaptable under extreme mobile network conditions, such as long periods of disconnection or low bandwidth and high latency. ODK-X replicates data to mobile devices, allowing the Framework to retain full functionality in disconnected environments.

The ODK-X suite offers various services to the user to keep their tool powerful and flexible. The most relevant points based on their official website [20], which offers fairly comprehensive documentation on the tool, are:

- **Synchronize bidirectional data**: A two-way synchronization protocol allows us to create data management applications with:
  - Monitoring surveys and data collection locations
  - Pre-filled forms for faster data collection
  - Data can be synchronized on all devices from the server
- **Offline data collection**: Allows users to collect data without an internet connection. The form data can be synchronized with the server when the user has Internet access.
- **Linked and embedded surveys**: ODX-X tools allow to open and edit other surveys with links to the original survey, and create a subform (nested) relationship between surveys or links relationships between data.
- **Data view on device**: Analyze and visualize entire data sets directly on the device via graphical, map and tabular views and filtered views.
- **User access control**: Control of privileges to view, modify and delete data for different users and groups.
- **Customizable survey feed and appearance**: Using basic web development (HTML, JavaScript, and CSS) to specify the layout of almost any screen seen by data collectors.

2.1 ODK-X: Classic collection process

The data collection phase is defined in the literature [21] as a means of acquiring raw data from one or more specific environments. Thus, we will follow the collection process proposed in [22] which we consider to be the classic process used by default and is defined in six steps.

- **Define the objectives of data collection**: Identify the CPGEs establishments of the students who obtained the best results in their 1st year of the engineering cycle.
- **Develop a list of questions of interest**: For example, the first question being "Did you spend your 1st year at this establishment?" aims to target our audience.
- **Establish data categories**: The fields to be completed and which are linked to the identification of the student are: Last Name, First Name, Gender, CIN (Identity Card Number), CNE (Student Card Number) and Registration Number. We consider that the student may not know his Registration Number which is linked only to the establishment, or even his CNE which is rarely used, hence the proposition of the CIN field as well. The Last Name and First Name fields also appear in the form out of habit.
- **The other fields related to this case are**: City CPGE, Name CPGE, Year of success in 1st year, Grade obtained in 1st year.
- **Design and test the data collection form**: The first section of the created form involves verification of the user: If it is actually among the targets of our investigation, whether he spent his first year at school. Thus, a first question is manda-

tory before completing the form: Did you spend your 1st year at the establishment? Accordingly, we must create the adequate lines in form under the Survey tab (see Table 1).

Table 1. Excel json code configuration file - display of the first screen of the Default Process.

Clause	begin screen			end screen
Type		note	select_one	
Values_list			yesno	
Name			Skip1Year	
Display.prompt.text			<center><b> Did you spend your 1st year at the school? </b></center>	
Display.prompt.image		img/school.png		
Required			TRUE	
Display.required_message.text			Please reply to:	

The answer to this question is mandatory to continue the process. The choices "Yes" or "No" are defined in the choice tab of the Default_Processus.xlsx form. An error message is then displayed, stating that this step is essential thanks to the mention "TRUE" in the "Required" column of the form under the Survey tab. On a negative response on the first screen of the survey, the user is considered off-target. Thus, an end of process screen will be displayed to the user.

Once this first step is verified and the user responds positively to the question, the second step is to define the fields to fill in to meet our objective. In this case, the type of each field represents the only check performed in the default process.

Regarding the form, no requirement has been defined, so no field will be required when entering. The choice of fields is made according to the following three aspects:

- The student's identifier to subsequently verify the accuracy of the information filled in during the analysis phase.
- The gender and year of success of the 1st year for an analysis by year and gender as expressed at the beginning.
- The grade for the first year at the establishment, city and name CPGE.

The type of fields is consistent with the nature of the information expected, the CNE and Registration Number are integer types. In order to minimize the number of errors entered, we have chosen to use a drop-down list for the following fields: Gender, CPGE Name, CPGE City and Year of passing the 1st year. While the Grade is a decimal field, the other fields like Last name, First name and CIN are of type text.

Gender accepts three values namely Female, Male and Other mentioned under the choices tab of the form. Changes to the choices sheet contain the response lists defined for our drop-down lists. The headers used correspond to:

- Choice_list_name: The group name for all the answers in a choice set.
- Data_value: The data value to select.

- Display.title.text: The text the user will see to select this value.

In the same way as the gender, we have filled in the Name CPGE and City CPGE with the complete list presented in official website. Our target being the students still present in the school, so the year of success in 1st year cannot be lower than 2015. When the user proceeds to the next step, he will find an end screen.

- **Collect and validate data:** This part will be described in section 5 (Results). It is about collecting data according to defined test cases.
- **Analyze data:** This part will be described in section 5 (Results). This is to discuss the results of the data collected.

2.2 Smart collection process

In this section, we will follow the Smart data collection process, which is made up of the seven sub-processes representing the steps [1].

Planning: The planning sub-process allows us to define the strategies to be followed, the requirements and to know our client better. This is a set of activities to be performed in order to define all the strategies relating to data collection.

- **Identify customer requirements**. The functional requirements raised are as follows:
  - Identify students who spent their 1st year at the institution
  - Know the Moroccan CPGEs of origin of the students who obtained the best grades in the 1st year at the institute
  - To be able to filter by genre
  - To be able to filter by year of obtaining in order to make an analysis for each promotion
- **Define the protocol's strategy**: Surveys [23] are a popular means of data collection because they are inexpensive and can provide a wide perspective. In this thesis, we have opted for the survey using a mobile device. Mobile data collection [24] is a method of compiling qualitative and quantitative information using a mobile device. This approach will allow us to increase the speed and accuracy of data collection, the efficiency of service delivery and the productivity of program staff.
- **Finally, to make our survey of better quality** and for validation purposes, we added a second method of data collection, which is “Textual or content analysis” for the reliability of the results provided and this thanks to the administrative file held by the school.
- **Define a search strategy**: Based on the CartOng article [13], a comparison of 26 tools used for mobile data collection allowed us to choose ODK-X as the tool that best meets our need.
- **Define an enrichment strategy**: The enrichment strategy will combine two approaches, the first being the survey that will be done regularly at the start of each

academic year among target students. The second approach is to complete the missing information if necessary based on the administrative file containing the grades and information necessary for each student, in order to have results reflecting reality and not just sampling.

- **Define a storage strategy**: The objective of this survey being to subsequently analyze the success keys of students from CPGEs, we will then keep track of data from the last 5 years only. Any data beyond 5 years will be destroyed to free up the space dedicated to storage.
- **Define an evaluation strategy**: The administrative file collects all of a student's information during their course. Based on the high rate of accuracy of the information present in this file, our assessment will be made by relying solely on the data provided by this file.
- **Validate strategies**: The confidence measure linked to the administrative file will allow us to estimate the quality and validity of the data. The goal is to have all the students who obtained the best 1st year grades at the institute for each academic year.

Protocol: The result of this planning allowed us to put in place the appropriate actions and strategies to create a form that best meets the need while being adapted to the process. Once the key objectives have been set, the Protocol sub-process describes the activities to be carried out to obtain the desired data.

- **Define inclusion/ exclusion criteria**
  - **Languages to include/ exclude**: We only include the same language used in the administrative file.
  - **Sources to include/ exclude**: Data collected through the forms ODK-X, refer to the first survey: 2nd year students who spent their 1st year at the institute, and 3rd year students who spent their 1st year at the institute.
  - **Beyond the 1st survey**, only 2nd year students who have spent their 1st year at the institute will be concerned by these surveys.

- **Define integration criteria**: We must unify values of the size of the field “Grade” obtained in the first year for its format in storage can be written with two decimal places.
- **If the user fills in at least one** of the “Last name” or “First name” fields, the value of the field must be masked when storing data to respect anonymity.
- **Choose methods to use**: Depending on the search strategy defined, the method assessed useful is the ODK-X software suite.
- **Define procedures to be used**: Following the defined protocol strategy, the procedure to be used will be a survey based on electronic forms that will be completed by the students of the institute who meet the pre-established criteria.
- **Define customer satisfaction rate**: To measure customer satisfaction, we will need to have a rate of over 70% of the population present to judge that our data collection is relevant.

Data management:

- **Receive data**

- Referential of procedures: In the first phase of the form we ask the user if he has already spent his 1st year at the institute. However, a footnote mentions that the answer to this question is only intended for Moroccan students, from a CPGE and who are still studying at the institute as 2nd and 3rd year students only. The objective of the note is to not waste students' time who have integrated the institute through a faculty who hasn't followed their first year at this institute for example.
- Development of data summary: When filling the form by a user, we notice the items described below and it's the disadvantage of forms in general: Incomplete forms, Duplicate forms, Forms with incorrect data
- Improve collection: To solve the problem of incomplete forms, we have put a number of management rules. Thus, one of the CIN, CNE or Registration Number fields must be filled in to successfully identify the student and make the necessary checks. An error message is displayed preventing going to the next step if the condition is not met. This view is possible thanks to the addition of the constrained column in the form under the Survey tab (see Table 2).

Table 2. Constraint on user identification

type	**name**	**display.prompt.text**	**constraint**	**display.constraint_message.text**
text	CIN	CIN	((data('RegistrationNumber') != null) || (data ('CIN') != null) || (data('CNE') != null))	One of the following fields must be specified: CNE, Registration Number or
integer	CNE	CNE	((data(' RegistrationNumber ') != null) || (data('CIN') != null) || (data('CNE') != null))	One of the following fields must be specified: CIN, Registration Number or
integer	RegistrationNumber	RegistrationNumber	((data(' RegistrationNumber ') != null) || (data('CIN') != null) || (data('CNE') != null))	One of the following fields must be specified: CIN, CNE or

Similarly CPGEs the Name field is also mandatory and implicitly CPGEs City, since it feeds the CPGEs name once a choice is selected from the list of available cities. This correspondence is done thanks to the choice_filter column which allows to create this link between the two columns. If the user does not enter the CPGE Name, an error message is displayed preventing him from going to the next step of the form. The choice to make this field mandatory is to know the list of CPGEs of students who obtained better grades in their 1st year. The display of this view is possible thanks to the addition of the constraint column in the form under Survey tab.

Finally, to meet the other requirements, we need to know the year of passing the 1st year as well as the grade obtained. Likewise, we've added an error message that displays if the condition is not met by adding the correct condition under the Survey tab of the form.

In order to improve the consistency of the data sent without reducing the performance of the form, we have opted to add a verification step after the user has filled in all the mandatory fields.

Thus, only one call will be made to the server for data verification in order to minimize the number of calls to the server and the response time. The data is checked first by CIN if filled in otherwise by CNE and in the event that none of these fields is filled in we check the data by Registration Number.

The verification is done on all non-empty fields even if the field is not mandatory, based on the administrative file present in the server. The call to this verification is made thanks to the column "calculation" in the form under the Survey tab with the value: VerifData (data ('Name'), data ('First name'), data ('CIN '), data (' CNE '), data (' Matricule '), data (' Gender1 '), data (' City_CPGE '), data (' Name_CPGE '), data (' year '), data (' Grade ')). The verif variable is used to define the order of verification when a user enters several identifiers. The identifiers are the CIN, the CNE and the Registration Number.

We then check the output value if it is false in this case the data is not compliant and the form cannot be finalized with an error message displayed. In this case, we are using the constrained column in the form under the Survey tab to display the error message. The choice not to mention the error explicitly is due to the fear of intentionally divulging information to an individual who is groping for the exact data with our help specifying the origin of the error.

- **Manage sources**: The list of CPGE names and the list of CPGE cities cannot be deduced from the administrative file because they can change from one year to another depending on the students admitted, for this we have put all the data in a csv file conforms to the official website. This file is read using queries built into the Smart_Processus.xlsx file under the queries tab. The call to this data is made in the survey tab of the same file by adding the value of the query_name field in the "values_list" column of the desired field.
- **Smart Data L0**: The result up to this point represents the valid data conforming to that present in the administrative file.
- **Create data**
  - **Criteria creation**: If the data is not sufficient to achieve customer satisfaction, that is at least 70% of the students answered the form, the missing data can be created internally until the desired threshold is reached.
  - The students concerned by this creation are student spent their 1st year at the institute and currently at their 2nd or 3rd year.
  - **Data validation**: The data will follow the same validation procedure as that presented when receiving the form from a student.
  - **Smart Data L0**: The result up to this point represents the valid data conforming to that present in the administrative file.
- **Integrate data**
  - **Sensitivity analysis**: Knowing well that our choice to keep the same graphic form as the one present in the classic approach, allows us to understand the influence of the missing data on the result. Reason why we send implicitly the value of the

Sector field from the administrative file, while sending the form to the server to complete the information that we judge important to analyse.

- **Statistical analysis**: In this case, we can analyze the number of students who filled in the CIN field, the CNE field and finally the Registration Number field, with the objective of knowing which of these fields can be defined for identification.
- We can also know the number of participants who did not spend their first year at the establishment and who still answered the questionnaire.
- **Analysis of data to be excluded**: In the context of anonymity, we can hide the first and last names of students, especially as this information does not add any value to us in our process. Thus, this check is added in the formulaFunctions.js file in JavaScript and called by the value securite_nom (data ('Name')) or securite_nom (data ('First name')) mentioned in the "calculation" column of the form under the Survey tab.
- **Analysis of conflicting results**: For this scenario, no data will be contradictory compared to another, since the value of the data is compared to the administrative file before completing the questionnaire.
- **Data comparison**: In our case, this step is not necessary. All data in the same column has the same type.
- **Integration:** We want the format of the Grade to be structured as follows 1X.XX with X € [0, 9]. The toPrecision method allows to format a number to a defined length, in our case we set it to 4. This format change will not impact data verification, but it will be implemented when sending the form to the server.
- In addition, we add a condition so that the "Grade" field is between 10 and 20 (data ('Grade')> = 10) && (data ('Grade') <20)). Note that the minimum passing grade may vary from year to year for another reason why we have set a score of 10 as the minimum threshold. If this condition is not met an error message is displayed.
- **Smart data L1:** The results obtained are valid data in accordance with those present in the administrative file while respecting the anonymity of the users by hiding their Name and First name if filled in by the student. All the values in the "Grade" field respect the same format. Finally, some necessary data is added when sending the form without an additional step by the user.

- **Assess**
  - **Evaluation of quality criteria**: The total coverage of customer needs is due to the good quality of the criteria established in the planning phase.
  - **Evaluation of data quality**: The quality of the data is respected and conforms to the requirements declared by the customer thanks to the conformity of the recorded values.
  - **Evaluation of integration criteria**: All the data are in the same format, so we can deduce that the integration criteria are well respected.

- **Synthesiz**e: Since this approach involves communication with an external server for data verification, the ODK-X architecture has undergone a slight improvement by using edge computing [25] in architecture.

Once the form is validated and saved to App Designer, synchronization with ODK Aggregate is the second step. We can then download the data from the server using the correct authentication. On a mobile device, data cannot be viewed or created unless the phone is synchronized with the server, thus allowing data to be downloaded.

We can thus start a new form, or follow the modification of a form that we have already initiated from ODK-X Survey or ODK-X Tables. The modification we have added to the process is the call to the external server which is only done at the last step of the form and this in order to lighten the process and only make the call when it is necessary. Knowing that a blocking step following three attempts by the same phone on the same form is envisaged to also limit the number of external calls.

Once the user completes the form, he can keep the changes only on his phone by updating OI File Manager, or submit it to the server. The administrator will be able to view the data sent to the server from his computer once synchronized.

Presentation of the results

- **Summarize**: Present the recorded data to the server.
- **Deduce the results from the data**:
  - Identify gaps: In this case, no deviation is identified by following this process.
- **Interpret the data**: The interpretation of the data makes it possible to identify the CPGEs from which the students who obtain the best grade in the 1st year of the engineering cycle come.
- **Define persistent problems**: A student can complete the form as many times as they want if all the information is correct.
- A user can try to fix errors on a form many times and this can be the cause of malware.
- **Identify improvements**: When submitting a form we can verify that the customer ID does not already exist in our database. Otherwise, we may display an error message stating that the user has already taken the survey.
- For the second issue raised we plan to set a countdown that will block the user from retrying if they exceed 3 attempts, until they are unblocked by an admin.

Enrichment

- **Analyze the need of enrichment**: The gender field is not mandatory, but we can do a gender analysis to have statistics by gender as well. Thus, we can add this value at the moment of sending a valid form based on the value present in the administrative file when it’s not entered.
- **Analyze the need of future enrichment**: We need enrichment data every year, once the 1st year students pass.

Visualization

This part will be described in the next section. It's about visualizing the results and discussing them.

3 Results and Discussions

The test cases allow us to put several scenarios that can be encountered when filling out the form. Then, we compare the results obtained for each sub-process based on these test cases (see **Fig. 1**).

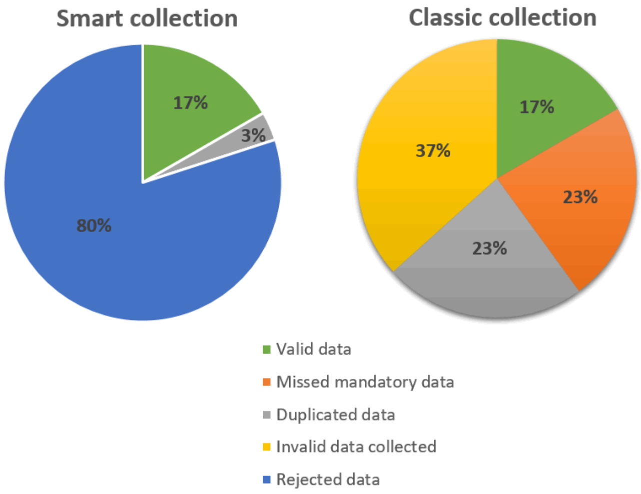

Fig. 1. Comparison measurements.

For example the first case that has the merit of being present is when the first question of the form is answered in negative way. We then begin the different cases with a positive answer to the first question on the form.

Unlike the classical approach results that returned all records, the result of the same test cases using the smart approach returned fewer result. Thus, the parameters used to measure the quality of data between the classic collection process and the Smart collection process:

- Number of data with empty mandatory field
- Number of duplicate data
- Number of invalid data (incorrect value)
- Data rejected

Using the classic collection process, we notice that all records are accepted even though in our case only 5 test cases are correct. In the other case, the Smart collection process will reject all the test cases with missing mandatory fields or invalid fields, but will only accept one case of duplicate data, which was noticed at the end of the process and noted as an improvement.

Thus, the result of our test using the Smart approach, allowed us to record only 6 results. Beside, to fully unfold our process, we need to launch a second test campaign given the low number of data collected. However, we have noticed that some cases of duplication can be generated due to a different use of the student ID. For this, we propose as an area of improvement for the next collection to add this verification.

4 Conclusion

In this paper, we have tested Smart data collection process via the ODK-X software suite and we have managed to identify the added value of this process compared to the one used by default during collection. To this end, we carried out the same test case according to the two processes. The first, which is by default, allowed all cases to be collected and sent to the server, although some data are worthless.

Upstream, our experiment raised the blocking points and the disadvantage of the workflow used by default to achieve a satisfactory result while meeting the customer's end goal. The user is guided so as not to provide incomplete or erroneous information and this in order to assemble clean and structured data to optimize analysis time and storage space. With architecture ODK-X we raised the added validation step does not change the overall system instead, it allows to refine the collection without impacting the performance of the tool.

One of the strengths of our process is the continuous improvement of the data and this thanks to the dedicated steps to check if the data meets the pre-established plan or not and if not, the action to be taken to cure it. Knowing that an evaluation of the methods used is also an asset in remedying a mistake in planning. The process is a way to organize the steps to follow while putting in place the means for continuous improvement and steps to a reflection on the choices to be made.

5 References

[1] I. Tikito and N. Souissi, "A Smart Process of Data Collect," Proceeding: Intelligent Systems and Computer Vision (ISCV), pp. 1-7, 2020b.

[2] E. Paradis, B. O'Brien, L. Nimmon, G. Bandiera and M. A. T. Martimianakis, "Design: selection of data collection methods," Journal of graduate medical education, vol. 8, no. 2, pp. 263-264, 2016.https://doi.org/10.4300/jgme-d-16-00098.1

[3] I. Tikito and N. Souissi, "Towards a systematic collect data process," International Journal of Big Data Intelligence, vol. 7, no. 2, pp. 72-84, 2020a. https://doi.org/10.1504/ijbdi.2020.107374

[4] . Tikito and N. Souissi, "Data Collect Requirements Model," Proceedings of the 2nd international Conference on Big Data, Cloud and Applications. ACM, p. 4, 2017.https://doi.org/10.1145/3090354.3090358

[5] I. Tikito and N. Souissi, "Methodology of Data Systematic Review: a step-by-step guide," Proceedings of the 3rd international Conference on Big Data, Cloud and Applications, 2018.

[6] A. Tella, "Electronic and paper-based data collection methods in library and information science research: A comparative analyses," New Library World, vol. 116, no. 9/10, pp. 588-609, 2015. https://doi.org/10.1108/nlw-12-2014-0138

[7] K. Avi, G. Nicholas, G. Thomas, K. John D and S. Mark J, "Validation relaxation: a quality assurance strategy for electronic data collection," Journal of medical Internet research, vol. 19, no. 8, p. e297, 2017. https://doi.org/10.2196/jmir.7813

[8] A. Flaxman, A. Stewart, J. Joseph, N. Alam, S. Alam, H. Chowdhury, M. Mooney, R. Rampatige, H. Remolador, D. Sanvictores, P. Serina, P. Streatfield, V. Tallo, C. Murray, B. Hernandez, A. Lopez and I. Riley, "Collecting verbal autopsies: improving and streamlining data collection processes using electronic tablets," Population health metrics, vol. 16, no. 1, p. 3, 2018. https://doi.org/10.1186/s12963-018-0161-9

[9] S.-W. Jun, M. Liu and S. Lee, "BlueDBM: Distributed Flash Storage for Big Data Analytics," ACM Transactions on Computer Systems (TOCS), vol. 34, no. 3, p. 7, 2016.

[10] H. Cai, B. Xu and L. Jiang, "IoT-based Big Data Storage Systems in Cloud Computing: Perspectives and Challenges," IEEE Internet of Things Journal, 2016. https://doi.org/10.1109/jiot.2016.2619369

[11] R. Kaur, I. Chana and J. Bhattacharya, "Data deduplication techniques for efficient cloud storage management: a systematic review," The Journal of Supercomputing, vol. 74, no. 5, pp. 2035-2085, 2018. https://doi.org/10.1007/s11227-017-2210-8

[12] M. Nayak and K. Narayan, "Strengths and weakness of online surveys," IOSR Journal of Humanities and Social Science, vol. 24, no. 5, pp. 31-38, 2019.

[13] CartOng, "Benchmarking of Mobile Data Collection Solutions," 26 Janvier 2017. [Online]. Available: https://blog.cartong.org/wordpress/wp-content/uploads/2017/08/Benchmarking_MDC_2017_CartONG_2.pdf. [Accessed 2019].

[14] P. L. Bokonda, K. Ouazzani-Touhami and N. Souissi, "A Practical Analysis of Mobile Data Collection Apps.," International Journal of Interactive Mobile Technologies, vol. 14, no. 13, pp. 4749-4754, 2020. https://doi.org/10.3991/ijim.v14i13.13483

[15] P. L. Bokonda, K. Ouazzani-Touhami and N. Souissi, "Open Data Kit: Mobile Data Collection Framework for Developing Countries," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 12, pp. 4749-4754, 2019. https://doi.org/10.35940/ijitee.l3583.1081219

[16] P. L. Bokonda, K. Ouazzani-Touhami and N. Souissi, "Mobile Data Collection Using Open Data Kit," in International Conference Europe Middle East \& North Africa Information Systems and Technologies to Support Learning, Springer, 2019, pp. 543-550. https://doi.org/10.1007/978-3-030-36778-7_60

[17] H. Carl, L. Adam, A. Yaw, T. Clint, B. Waylon and B. Gaetano, "Open data kit: tools to build information services for developing regions," Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development. ACM, p. 18, 2010. https://doi.org/10.1145/2369220.2369236

[18] Y. Anokwa, C. Hartung, W. Brunette, G. Borriello and A. Lerer, "Open-source data collection in the developing world," Computer, vol. 42, no. 10, pp. 97-99, 2009. https://doi.org/10.1109/mc.2009.328

[19] B. Waylon, S. Samuel, S. Mitchell, L. Clarice, B. Jeffrey and A. Richard, "Open Data Kit 2.0: A services-based application framework for disconnected data management," in Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services. ACM, 2017. https://doi.org/10.1145/3081333.3081365

[20] odk-x, "opendatakit," 2020. [Online]. Available: https://docs.opendatakit. org/odk-x/. [Accessed 2020].

[21] I. Tikito, M. El Arass and N. Souissi, "Meta-Analysis of Data Collect Methods," Journal of Computer Science, vol. 15, no. 8, pp. 1184-1194, 2019. https://doi.org/10.3844/jcssp.2019.1184.1194

[22] V. R. Basili and D. M. Weiss, "A methodology for collecting valid software engineering data," IEEE Transactions on software engineering, no. 6, pp. 728-738, 1984. https://doi.org/10.1109/tse.1984.5010301

[23] S. M. Kabir, Basic Guidelines for Research: An Introductory Approach for All Disciplines, First ed., B. Z. Publication, Ed., Bangladesh, 2016, pp. 201-275.

[24] A. Samaddar, A. Ajay, A. Keil, A. Rai, S. Sharma, S. Pal, A. Arora, S. Marwaha, S. Islam, A. Gupta and R. K. Paul, "Open data kit for diagnostic crop production survey at landscape level in India," International Maize and Wheat Improvement Center (CIMMYT), pp. 11-17, 2019.

[25] H. Li, K. Ota and M. Dong, "Learning IoT in edge: Deep learning for the Internet of Things with edge computing," IEEE network, vol. 32, no. 1, pp. 96-101, 2018. https://doi.org/10.1109/mnet.2018.1700202

6 Authors

Iman Tikito has more than 8 years of international experience as Business Analyst and Engagement manager, working for a multinational company. She holds a double Master Degree in IT Applied to Offshore Development from the University of Mohammed V at Morocco, a master degree in Offshore Development of Information Systems from University of Bretagne Occidental at France. She’s currently pursuing her Ph.D. in Science and technology for the engineer at Mohammed V University in Rabat, EMI-SIWEB Team, Rabat, Morocco. Email: iman.tikito@gmail.com

Nissrine Souissi is a full Professor at Systems Engineering and Digital Transformation Laboratory (LISTD), SSDT Team, Computer Science Department, MINES-RABAT School (ENSMR), Morocco. She obtained her PhD in computer science from the University Paris-Est Creteil (UPEC) in 2006, France and an Engineer degree from Mohammadia School of Engineers (EMI) in 2001, Morocco. Her research interests include process engineering, business process management, digital transformation and data engineering. Email: souissi@enim.ac.ma

Article submitted 2021-03-28. Resubmitted 2021-04-29. Final acceptance 2021-04-30. Final version published as submitted by the authors.