CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Collection Session 1

10-12 October 2018, Geneva, Switzerland 10 October 2018

CAPI-STIS: Integrated Digitalized Data Collection Software System for Official Statistics Survey

Takdir (Politeknik Statistika STIS, Statistics Indonesia)

takdir@stis.ac.id

Abstract

Data collection is the critical part of the whole survey or census phase which highly affect the data quality. Computer-Assisted Personal Interviewing (CAPI) technology combines enumeration and data entry into one unified process. A survey based on CAPI involves many stakeholders, such as interviewer, supervisor, and questionnaire designer. Besides software and hardware, CAPI implementation in a survey also has to be well organized and managed properly. An occurrence of failure in the use of digital questionnaire without planned backup will result in harm, e.g. invalid data and termination in data collection or data processing.

We studied and proposed essential rules for managing and organizing survey when CAPI implemented, especially in large scale official statistics survey. Various problems and findings related to CAPI implementation were gathered from previous CAPI applications and pilot projects established by Statistics Indonesia (BPS) and/or field studies in STIS, Institute of Statistics, Indonesia. This study analyzed and elaborated that information to determine the appropriate organization and management of CAPI-based survey. The result is a best practice for conducting CAPI-based survey in order to deal with problems arisen or at least to reduce the harmful effects.

Furthermore, we introduced a CAPI software, namely CAPI-STIS, including the development roadmap considering our investigation to the existing available software for CAPI. It was developed from the famous opensource CAPI software Open Data Kit (ODK) with major modification adjusting with official statistics survey characteristics. It is multimode data collection that can be used with mobile phone, and web interface, and additionally integrated with survey progress and raw data monitoring. Our system has been tested and implemented for numerous important survey such as Indonesia Socio Economic Survey (SUSENAS), poverty survey, and horticulture survey. CAPI-STIS is expected to become a reference in developing software for critical survey, particularly for official statistics.

CAPI-STIS: Integrated Digitalized Data Collection Software System for Official Statistics Survey

Takdir
takdir@stis.ac.id

Politeknik Statistika STIS, Statistics Indonesia
Kampus Politeknik Statistika STIS
Jl. Otto Iskandardinata No. 64C
Jakarta 13330
Telp. (021) 8191437, 8508812
Fax. (021) 8197577
http://stis.ac.id

1. Executive Summary

Data collection is the critical part of the whole survey or census phase which highly affect the data quality. Computer-Assisted Personal Interviewing (CAPI) technology combines enumeration and data entry into one unified process. A survey based on CAPI involves many stakeholders, such as interviewer, supervisor, and questionnaire designer. Besides software and hardware, CAPI implementation in a survey also has to be well organized and managed properly. An occurrence of failure in the use of digital questionnaire without planned backup will result in harm, e.g. invalid data and termination in data collection or data processing.

We studied and proposed essential rules for managing and organizing survey when CAPI implemented, especially in large scale official statistics survey. Various problems and findings related to CAPI implementation were gathered from previous CAPI applications and pilot projects established by Statistics Indonesia (BPS) and/or field studies in STIS, Institute of Statistics, Indonesia. This study analyzed and elaborated that information to determine the appropriate organization and management of CAPI-based survey. The result is a best practice for conducting CAPI-based survey in order to deal with problems arisen or at least to reduce the harmful effects.

Furthermore, we introduced a CAPI software, namely CAPI-STIS, including the development roadmap considering our investigation to the existing available software for CAPI. It was developed from the famous open source CAPI software Open Data Kit (ODK) with major modification adjusting with official statistics survey characteristics. It is multimode data collection that can be used with mobile phone, and web interface, and additionally integrated with survey progress and raw data monitoring. Our system has been tested and implemented for numerous important surveys such as Indonesia Socio Economic Survey (SUSENAS), poverty survey, and horticulture survey. CAPI-STIS is expected to become a reference in developing software for critical survey, particularly for official statistics.

Regarding the objective of the workshop where this paper submitted, to identify innovative ways and best practices in statistical data collection, we organized the structure of the explanation slightly different from academic paper. This paper contains best practices rather than evidence of our experiments and observations.

2. Politeknik Statistika STIS

In 1964, Statistics Indonesia, the National Statistical Office (NSO) of Indonesia, (in Bahasa: Badan Pusat Statistik (BPS)), established a vocational higher education in statistics, namely AIS, stand for Akademi Ilmu Statistik. The lectures were from the United Nations. In 1997, AIS became STIS, stand for Sekolah Tinggi Ilmu Statistik, and increased the level of graduate competency from Diploma III to Diploma IV, which is equivalent with undergraduate or Bachelor degree. STIS also introduced new study program, Computational Statistics (or Statistical Computing) to fulfill the demand of human resources with knowledge of statistics and skill in computer programming for improved data processing in Statistics Indonesia. Just in 2017, STIS transformed to Polytechnic, Politeknik Statistika STIS, which has equal organizational structure with university as well as the opportunity to establish graduate degree program in applied science. Computational Statistics Department encourages their lecturers and students to take advantage of cutting-edge technologies and innovations to improve official statistics process and data quality. In last 3 years, STIS have utilized CAPI as the main data collection tool in annual student's field study. The CAPI software and methodology are then adopted by some surveys in Statistics Indonesia, such as socio economic and horticulture survey.

3. CAPI Implementation in Official Statistics

CAPI introduces some advantages in operation of survey, including speed (time), data quality, and operational cost. However, a special case is faced by NSO when adopting CAPI as the data collection tools for official statistics. Private and public sector organizations differ in structure, in the nature of surveys they conduct, and in the types of data they collect (Computer Assisted Survey Information Collection, Reginald P. Baker, 1998). The characteristics of official statistics, which is managed by government agencies, are:

- Large amount of sample sizes,
- Typically take a long time in overall survey process,
- High complexity in terms of variables collected, questions, organizations, and administration,
- Involves many different level of employees,
- Repetitive survey (annually, quarterly, monthly, etc.).

The existence of those particular features should be considered properly. The publicly available CAPI tools and methodologies need some adjustment, especially to deal with the survey complexity.

There are many software vendors and developers provide CAPI application with generic and adjustable functionalities, but the complex structure of official statistics, which involves hierarchical user privileges and multilevel validations, makes the acquisition complicated.

In the other hand, in some aspects, management and survey organization of CAPI are different from Paper and Pencil Interviewing (PAPI). Designing and testing of applications and questionnaires require greater effort. In other words, many thoughts are focused on accommodating the needs of the interviewer (Blackshaw et al., 1990). For massive adoption, restructuring the NSO and reengineering its business process should be performed.

4. Gradual migration from PAPI to CAPI

It is not trivial task to convince the senior survey manager or statistician to fully migrate their conventional paper-based survey into CAPI-based survey. In 2015, we compared the performance of CAPI implementations against PAPI in our annual students' field study in West Java province. There are 108 interviewers using CAPI and 304 using PAPI. 1,755 respondents are enumerated using CAPI or 21.49 percent of the sample sizes. The resulting survey timetable from field enumeration until the data is ready to analyze is shown in Table 1.

Table 1. Evaluation results

Activities	CAPI	PAPI
Field enumeration	7 days	7 days
Batching, Editing, Coding	-	16 days
Data entry	-	3 days
Ask supervisor to finalize data	3 days	-
Data cleaning	2 days	-
Total Time	**12 days**	**26 days**

Three interviewers are coordinated by a supervisor. The supervisors are also responsible to review the recorded data sent from their interviewers' devices. The supervisor can decide to return invalid data, provided with comments, to his interviewers if data anomalies were found. Once data finalized by supervisor, it cannot be modified from supervisor and interviewer devices. Batching, Editing, and Coding (for open questions) activities in PAPI include secondary check and correction of filled questionnaires, therefore invalid data can be minimalized in data entry. Additional data cleaning is required in CAPI as the substitution of that process. Nevertheless, CAPI shown significant time efficiency with the disappearance of Batching, Editing, and Coding.

In the next year filed studies, 2016, we increased the proportion of CAPI because the previous trial demonstrated good result without significant trouble. There were 228 interviewers using CAPI or 49 percent of all involved interviewers. Those CAPI interviewers were responsible for 3,406 respondents or 60 percent of the total respondents in Nusa Tenggara Barat province. We implemented Open Data Kit (ODK), which supports dynamic template-based digital questionnaire. It significantly reduce software development effort and responsive to questionnaire changes. Small amount of blind spots were reported by enumerators which cause the collected data could not be sent to the central server soon after the interview. The CAPI software has the ability to work in offline mode and data will be sent when internet connection available.

In the following years, our annual field studies was fully CAPI-based survey. The students become familiar with the use of digital questionnaire. Statistics Indonesia provided us 200 tablet PC, 40 percent of required devices, and the remainders used their own smartphone, known as *Bring Your Own Device (BYOD)*, in 2017. CAPI mobile application was equipped by both sample frame listing and the actual enumeration questionnaire. Some new additional features were introduced, such as automatic sample selection using predefined formula, and helpdesk messaging system. Latest, in 2018, BYOD becomes the main approach in our fully CAPI-based survey as the students tend to use their own smartphone instead of supplied devices. Real-time monitoring of important aggregate variables was provided and there were some improvements of helpdesk messaging system which can serve massive report from the respondents to be followed up by 24-hours live support.

From our experience, gradual, step-by-step migration from traditional PAPI to fully CAPI with continuous improvement can minimize the disruptive effect, increase confidence level, and prevent technology shock of all participating teams in survey.

5. Organization and Data Flow

Actors

Actors involved in CAPI implementation have their respective roles which probably different from those done in PAPI.

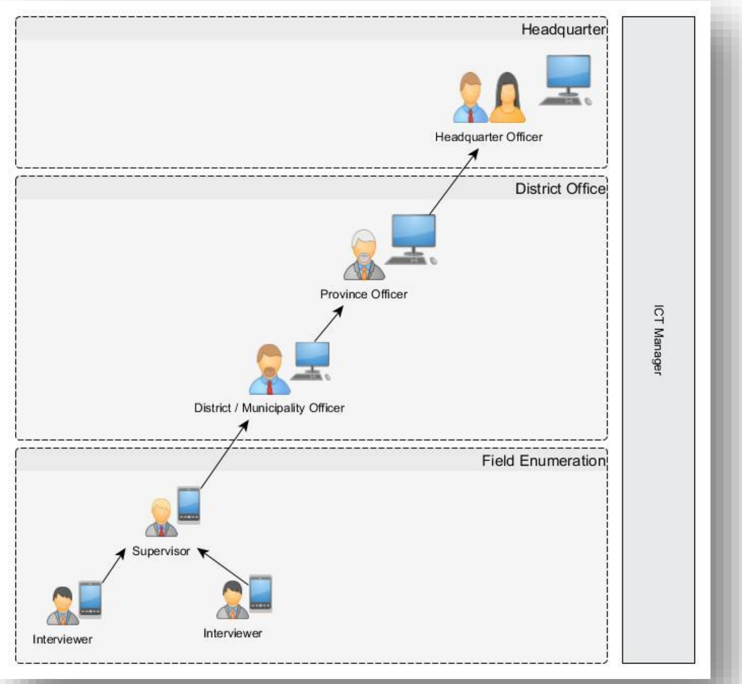

Figure 1. Organizational structure and data flow

1) Enumerator/Interviewer
Workloads that experience significant changes are at the level of the interviewer / enumerator. Besides conducting interviews with respondents, they have additional task as data entry. Based on the hierarchy of knowledge transfer (from headquarter until enumerators), they have the least understanding of the concept and technical definition of the survey carried out. Therefore, using CAPI, interviewer should have higher level of qualification and extra training, especially in using mobile devices. The infrastructures, including software and hardware, also should be conditioned to minimized difficulties of interviewers in filling digital questionnaire which can cause biased data recorded. The absence of paper documents that could become a reference if there are doubts on digital version data is an important consideration.
2) Field Supervisor
In PAPI, field supervisor checks the validity of data in questionnaire filled by enumerators because there is no control to limit data writing on paper. With the application of CAPI, the examination of the correctness of the contents is carried out automatically by CAPI application. Only valid entries can be entered into digital questionnaire.
There are 3 types of validation rule in CAPI:
   1. *No constraint*. All values can be entered,
   2. *Allowed*. Out of range value is allowed to be entered preceded by a warning message,
   3. *Restricted*. Out of range value is not allowed to be entered.

The complexity of rules is the main consideration to be applied in digital questionnaire. Complexity in this case is the existence of chaining rules in a field which has a relevancy to the value of other entries, for example the relationship between age and last education. Under this circumstance, the violation of a rules can be caused by invalid entries in the previous fields. The application of *restricted* rule in this case will make it difficult for interviewer to detect where the root cause of invalid entries. It will hamper the enumeration and cause incorrect entries due to probing entries to pass the rule checking.

Other case that does not allow the *restricted* rule to be applied is when there is a possibility of out of range value. For instance, if the rule set that the maximum normal price of rice is 15,000 IDR/kg, it is generally known and can be predicted that there is a possibility that the price of rice is above 15,000 IDR/kg in certain areas with extreme conditions. In those two cases above, *allowed* rule is the proper choice.

In CAPI, field supervisor plays a role in checking entries that have *no constraint* and *allowed* rules. Entries that violate the *allowed* rules can be automatically marked / highlighted by the application to navigate him/her to the fields that need to be corrected. If field supervisor find invalid entries, he/she can send notification to enumerator to re-check that entries.

3) Field Coordinator

In some surveys, e.g. the 2017 Potential Horticulture Survey (SHOPI), there is a field coordinator who execute samples generation and allocate them to enumerators. In another survey, that task is carried out by field supervisor. Some subject matters consider that the field coordinator has a strategic position due to his responsibility to avoid manipulation in the selection of samples, therefore it would need to be delegated to a special officer other than field supervisor.

4) District/Municipality Officer

Entries that was approved by field supervisor could ideally be accessed by district/ municipality officers. In the initial stage, the officer checks the statistics of data at the district level, then the sub-district, until the smallest grid of sampling area, namely census block. Checking is performed by looking at the aggregate value and making comparisons with the previous existing data to detect uncommon changes in data or trends. If required, the checking can be performed until the questionnaire entries level, and then followed up by contacting the field supervisor. Our observation shows that ideally district officer can revise invalid data without having to return the data to field supervisor or interviewers level to streamline the process but confirmation to field supervisor is required.

The common troubleshooting of devices and configurations was handled by municipality IT staff.

5) Province Officer

Equivalent to municipality officer, Province officers are responsible to ensure all data submitted by municipalities satisfy rational aggregate measurement in province level. Province performs managerial and organizational support to municipalities by bridging them to headquarter. IT department in province govern ICT devices including the procurement, registration, and distribution, procurement.

6) Headquarter

Headquarter compiles raw data from each province for further analysis. Advanced revalidation and data cleansing are carried out to produce final microdata.

6. Software Components

The software components can be categorized into main and supporting components. Main components are intended for the core survey activities involving the collection system and associated tasks, from data collection until clean data produced. It includes data collection tools in form of CAPI mobile applications and web based data entry for multimode survey purpose, data cleaning which contains algorithm and manual inspection of the recorded data, and reporting tools to provide the subject matter aggregate statistics view of the collected raw data. Supporting components play a role in equipping the main components with required entities for the operational of survey, such as the database of employees' attributes including their education background and track records, helpdesk systems as the main channel to communicate with online helpdesk support, and required specific custom components that can accelerate the operational like unit price converter from local measurement unit to standard measurement unit.

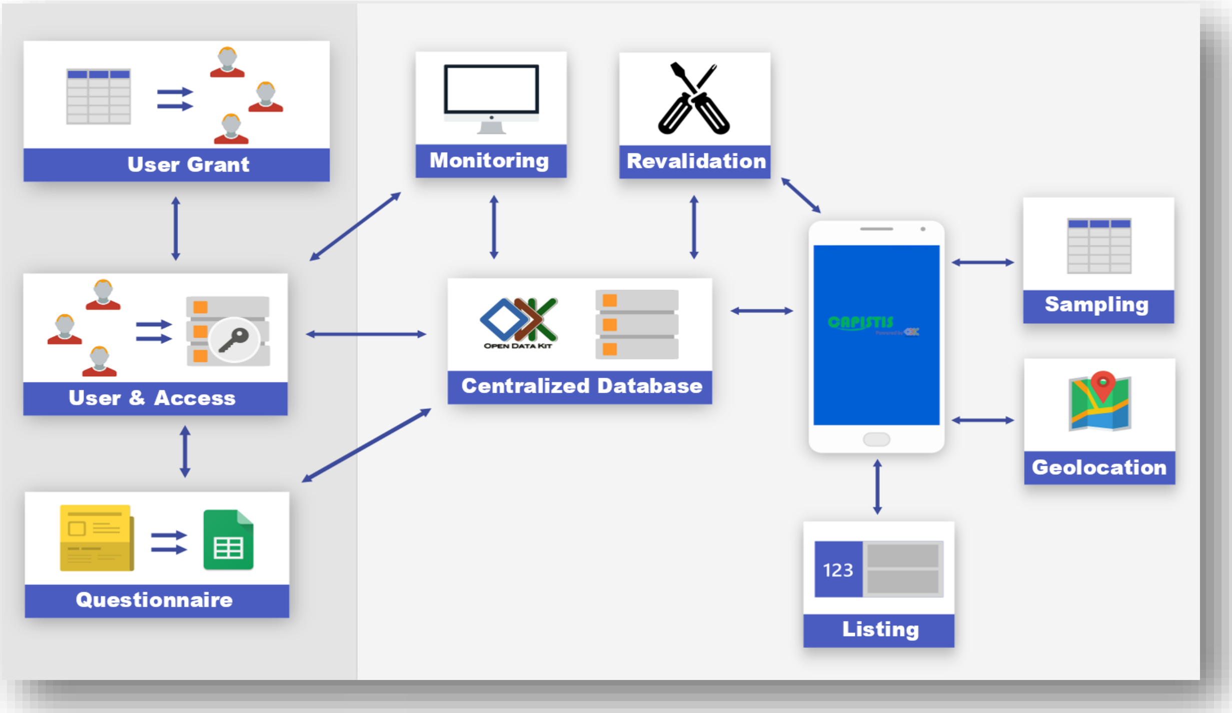

Figure 2. CAPI-STS systems' modules

Our CAPI-STIS utilizes open source software which are integrated into a survey environment system. These are includes:

a) Main data capturing tools. Forked from open source software, Open Data Kit (ODK) with major modifications, such as:
   - UX: Material Design. "Design is not added value, design is value", -- Gui Bonsiepe –
   - Search and navigation
   - Roster form questions
   - Multilevel user privileges
   - Real-time push notifications

b) Sample frame listing features to support sampling frame registration as commonly performed by NSO. The development includes:
   - Initial data upload
   - Inter-questionnaire dependency and validation
   - UX: tabular form

c) Multimode support. Providing another mode of data collections, e.g. web-based self-enumeration and traditional paper-based questionnaire, are integrated with the CAPI system in terms of data storage, data structure, and validation rules. This features enable the questionnaire structure and validations are shared among different collection modes. We thanks to the Enketo ® which provides offline web version of xForm, which has similar with the ODK questionnaire structure. We use it and modify some points to make it smoothly integrated with our developed systems.

d) Geolocation with offline support with MapBox framework.

All those features are linked each other with centralized control.

7. Conclusions and Remark

This paper briefly explain our current effort to modernize the data collection system, especially the use of information technology to foster the development of official statistics. There are many other aspects and challenges that we face either technical or management aspects. We proposed our approaches that has been implemented and tested by our institutions. We are still working to expand our system to meet the requirement of official statistics. At the end of the day, we will package this system to be used by other organizations, especially by government agencies, in order to improve their data collection efforts following the official statistics standard and platform. As the expected result, their collected data can be compared and mapped each other effortlessly.

8. Acknowledgement

CAPI-STIS have been studied and developed in many modules. The study involved several subject matters, especially Social Welfare Directorate and Statistical Information Systems Directorate of Statistics Indonesia. The milestone developers are M. Tohir, Budi Setiawan Akkas, I Gede Ananda N., M. Kaddafi and Rahadi Jalu Yoga Utama under the author supervision. M. Ari Aggorowati was studying on managerial and organizational support aspects of CAPI. Annual field study in Politeknik Statistika STIS uses our system and give a lot of inputs for future improvement.