

Large Scale Digital Data Collection in Developing Countries: Is The Time Right?

Mahier Hattas¹, Marius Cronje², Oliver Berard³

¹Statistics South Africa, e-mail: mahierh@statssa.gov.za;

²Statistics South Africa, email : mariusc@statssa.gov.za

³Easy Capture, e-mail: oliver@easycapture.org

Abstract

Statistics South Africa (Stats SA) is a governmental organization, mandated to publish official statistics in South Africa. Stats SA has recognized the value of digital data capture and thus has, with Easy Capture, collaborated in this paper to explore the viability of digital data capture (DDC) in developing countries for large scale household surveys. The methods for DDC were to use the Easy Capture system. Easy Captures past survey operations in combination with the Quarterly Labor Force Survey (QLFS) pilot formed the methodological design. The results with regard to the digital interface and skip pattern effectiveness; data exporting into pre-determined data template; enumerator adaption to digital system; in-field response to device usage; speed of data transfer from device to server; all contributed positively to the viability of DDC. The condition of the data results was the only area where there were errors. However, these were seen to be more related to user error and training constraints rather than technological issues. From a macro technological perspective and micro systems perspective it can be conclude that the time is right for DDC. From a National Statistical Organization (NSO) perspective the benefits that DDC can bring are highly significant and widespread. The time is right technologically, but what is needed is commitment from NSO's to build systems with private enterprise in highly collaborative efforts. Only then will the technology serve statistics as it should.

Keywords: Digital Data Capture (digital devices used in the face to face interview process to capture data), Household Surveys (a survey that collects information about the occupants of a household from their premises), Devices (A portable computing device capable of transmitting data)

1. Purpose

The purpose of this paper is to investigate the viability of digital data collection (DDC) in developing countries for large scale household surveys.

2. Background

Accurate and timely information is needed for decisions on the complex economy and social problems that confront a country. Large-scale household surveys have

traditionally been a very good source, throughout the world, for providing information on population, health, education, household income and expenditure, employment, and other critical areas of study. Large-scale surveys form part of three major sources for social and demographic data. These are surveys, censuses and administrative data (United Nations, 2005). The reliance on household surveys is even more important for developing countries that do not have a very sophisticated registration system or very reliable administrative records. Censuses are massive and expensive undertakings that some countries cannot afford. The reliance on household and related surveys means that the quality and timing of these surveys is of the utmost importance. There are various endeavors throughout the world to try and improve on these aspects.

The flexibility of household surveys makes them an excellent choice for meeting data users' needs for statistical information, which otherwise would not be available and insufficient (UN, 2005). Throughout the world large scale surveys are conducted through personal interviews using paper instruments as the tool for collection. These paper instruments are then sent to a central location to be captured onto a database using a basic capturing application. More recently, these paper instruments are scanned using a variety of character and image recognition engines. The scanned information is converted to usable data and is then edited prior to population of the central database. Questionnaires that are unscannable would either be transcribed or recaptured using a capturing application to the central database.

Information technology development during the past two decades has forced a rethink of methods and techniques used for survey processing. It is eminent that the next phase of conducting large scale household surveys should utilize handheld digital devices for data collection.

The United Nations (2005:210) states that:

“... although the technology has been available for many years, very little has been done to seriously apply this strategy to complex surveys in developing countries.”

They also identified potential problems related this method.

It is some of the aspects of the DDC process that is being investigated further in this paper. The paper explains in short the processes that takes place, it looks at a few tests conducted using the methodology and it discusses the results of the process from a technological perspective as well as a statistical producer perspective.

3. Methods

In the outlining of the methods, the Easy Capture DDC system will be expounded.

3.1 Device used: HTC P3300

3.2 Primary (minimum) requirements for DDC system to run on a device:

- Windows Mobile Operating Systems

- Touch screen interface
- GPRS capability

The DDC software runs off the above minimum requirements.

3.3 Cellular network used for data transfer: Vodacom South Africa

3.4 The process:

- Customer gives data template and form design (with skip patterns) to Easy Capture
- Easy Capture creates the digital version of the form and the data export template.
- Easy Capture checks and tests the survey and the export with the customer. Adjustments are made if necessary
- The completed and approved digital survey is then synchronized onto the devices via GPRS connection.
- User names are then created for the enumerators. And the enumerator's specific surveys are assigned to their user name. Enumerators are then trained on the devices.
- Enumerators conduct the surveys in the field on the devices
- Completed surveys are sent directly after completion via GPRS to the Easy Capture server.
- The data is then extracted off the server, exported and presented in the data template for editing.

3.5 Exploratory Tests:

This method/system of DDC collection has been used on survey operations locally in South Africa.

This paper will be drawing on the results from the QLFS pilot and the experience that Easy Capture has gained in DDC through other survey projects which in the light of this paper can be seen as exploratory tests.

For the QLFS pilot a total of 40 households were selected in 4 different enumeration areas (or Primary Sampling Units). 2 geographical area types, viz. "urban formal" and "urban informal" will constitute 20 interviews each respectively. 4 survey officers will administer 10 interviews. See table below:

4. Results

The DDC method that has been outlined above has been used in local exploratory tests. In testing for viability we observed the following results:

4.1 Digital interface and skip pattern effectiveness:

There are certain constraints when working with digital devices: The size of the screen and the limitations of the question design on the software. The main objective in the conversion from a paper form to digital one is not to lose the essence of the question being asked. The question being asked must draw out the same response as if it was being asked off a paper form.

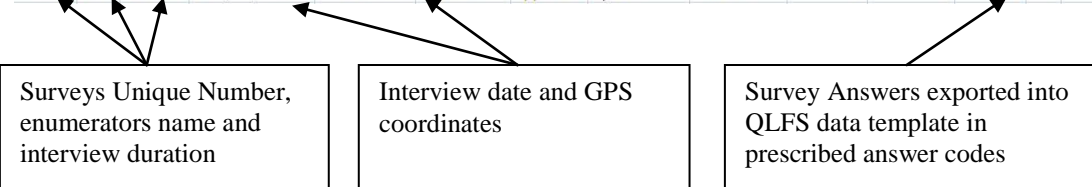
The QLFS was able to be satisfactorily converted without any major changes to the design of the questionnaire. The main change was in the QLFS Cover Page. Here 'response codes' and planned 'next visits' were structured slightly differently. Certain numeric questions that required the totaling of 'number of hours worked' were removed from view as the software was programmed to calculate them automatically. The automated skip patterns were found to work highly effectively, as all possible paths through the survey logic were successfully worked through by experienced QLFS enumerators. The answers received from the respondents, prompted from the digital interface, were consistent with those associated with the paper method.

4.2 Data exporting into pre-determined data template:

A data template was created that correlated with the QLFS data template. The question and answer codes from the original data template were created and embedded within the digital questionnaire. Thus, as interviews were completed and sent to the server they were exported into spread sheets that represented the original QLFS spreadsheets that were prepared for the paper system.

This image is a snap-shot of the QLFS spreadsheet that the DDC completed data is exported into.

SbjNum	Srvr	Duration	Upload Date	Latitude	Longitude	Status	Q11Name	Q11Surname	Q12Nights	Q13Gender	Date test	Q_7	Q14Day
30499	E2	00:33:39	2009/01/16 14:43	-33.90681333	18.52132167	Approved	Lesley	Williams	1	2	-1	-1	11
30500	E2	00:24:57	2009/01/16 14:43	-33.90681333	18.52132167	Approved	Zhaunette	Williams	1	2	-1	-1	7
30506	E2	01:21:01	2009/01/16 15:50	-33.929665	18.60011	Approved	Doris	Adams	1	2	-1	-1	7
30508	e4	00:11:38	2009/01/16 17:23	-34.11590333	18.85394167	Approved	Mzomhle	Thathoba	1	1	-1	-1	11
30510	e4	00:09:03	2009/01/16 17:58	-34.11604	18.85404167	Approved	Zingiswa	Msizi	1	2	-1	-1	20
30517	E2	00:08:46	2009/01/17 09:35	-33.938665	18.455145	Approved	Ayden	Williams	1	2	-1	-1	16



4.3 Enumerator adaption to digital system.

The QLFS enumerators were trained on the devices over two sessions on separate days. They took to the device and the interface quickly and after the training independently made their way through the digital questionnaires.

There were some enumerators who needed more attention than others on the devices. However, after the training sessions all enumerators were proficient in the use of the device in conducting surveys.

QLFS pilot in action: An enumerator from StatsSA interviews a household respondent



4.4 In field response to device usage:

The devices were not damaged, lost or stolen during field operations. There were concerns that the use of high end devices in poorer informal areas may attract unwanted attention. However, there were no problems of this nature reported.

4.5 Speed of Data transfer from device to server:

The Easy Capture DDC system is designed to send the ‘completed’ surveys to the server as soon as they have been completed. Thus, as an enumerator finishes an interview it gets sent. The speed that it took for the QLFS completed interviews to be sent to the server was between 0.2 to 2 minutes. The DDC system also allowed for new interviews to be started while completed interviews were being uploaded to the server.

4.6 Condition of data results:

The results from the QLFS pilot were populated as mentioned above into the data spreadsheets. The controlled skip patterns and survey logic meant that there were few errors from the interview process with the individual householders. Enumerators were able to open up the surveys that they had been completed on their devices and go through the survey again checking their answers. This meant that the enumerators could at least check their work before they sent the data off to the server. The one area where there were a few errors was in the recording of the ‘response codes’ and the scheduling of ‘next visits’. Some enumerators sent in Cover Pages that did not have response codes, or failed to enter a date for the next visit if the response code required it.

5. Discussion

5.1 A technical perspective:

5.1.1 Macro Conditions:

The advancement of cell phone technology, decreasing prices and the development of cellular network infrastructure has meant that the viability of digital data collection is at hand. The exponential increase in processing capabilities (according to Moore's Law, computer processing capacity doubles every 18 months) of handheld devices has led to powerful software applications being developed for these devices. One of these applications is form building software for data collection. This software, combined with ever decreasing prices for high capacity devices, makes DDC increasingly attractive. Rapidly advancing cellular infrastructure is providing wide spread network coverage (GPRS/GSM) for data transfer in South Africa and other developing countries. Thus, any data that is collected from handheld devices is able to be sent via GPRS to a specified server at relatively low cost.

5.1.2 Micro systems perspective:

Digital interface, Software robustness and flexibility:

The software used for the QLFS and other projects was able to successfully transform a paper based questionnaire into a digital version. The differences in design were minimal. The flexibility of the software in form design meant that no new questions had to be specially re-formulated. The software was robust in that there were no reported errors or freezing while in the field. All surveys when completed were successfully sent to the server.

The software being used was intuitive and simple to use. The software although being powerful and capable of conducting long complex questionnaires, is simple from a user perspective. The logical skip patterns and numeric calculations all happen behind the visible interface. This makes the training on the devices and the ability of the field worker to adapt to the new system favorable.

The built-in survey logic meant the enumerator had to spend less time thinking about skip patterns and could thus focus more on engaging with the respondent. This contributes to a stronger bond between the enumerator and respondent which should translate into more truthful and reliable data.

Enumerator adaption to digital system:

Cell phone penetration in developing economies can be seen as a major contributing factor to the ability of enumerators to adapt the new technology. The spread of cell phone technology to developing economies has meant that the broader population has been able to move up the technology curve. Thus, the adoption of devices by enumerators has been aided by learnt behavior and an implicit understanding of cell phone technology.

In the training this implicit understanding was drawn upon, helping the enumerator see the device as an adaptation of something they are already comfortable with, rather than a completely new concept. This increased their confidence in dealing with the new technology. Technology will only work well if the people using it believe in it, understand it, and know how to use it. The above results suggest that 'buy in' from enumerators for such a system would not be difficult to achieve given the correct training.

Speed of results:

The results are able to travel over GPRS to the server at very high speeds. The data is then ready for exporting into spread sheets immediately. This leads to the ability to have close to real time data quality assurance. The consequence of having data sent over GPRS means that most of the need for data processing is eliminated. There is no couriering of forms, no scanning of paper forms, no verification of scanned forms and no double data entry. New systems for data editing and quality control will be built around the new digital system as data processing cannot be entirely removed. This is because of the need of ensuring that there is still strict control over the final data presented for analysis. The cost savings and time savings as result of the elimination of paper data processing systems is one of the major benefits of DDC.

Data Export and Data Quality:

The surveys that were completed and exported were without error except for the Cover Page errors in 'response codes' and 'next visits' as mentioned above in the results section. The nature of these errors was due to a lack of training. There was not a problem with logic of the devices, rather a user error with regards to the process of completing and sending of forms with the correct 'response codes' and 'next visit' dates.

The data that was collected for the individual householders was satisfactorily exported. Data was received from all the completed interviews and exported accurately into the pre scribed data template. Thus, from a system perspective the data travelled smoothly into the correct fields.

5.2 An NSO Perspective:

From a National Statistical Office perspective it is important to evaluate the benefits that using this technology can bring to the organization. To do this from a quality point of view we will be using the *South African Statistical Quality Assessment Framework (SASQAF) (2008)*. This framework was developed using the generic framework of the Data Quality Assessment Framework (2003) of the International Monetary Fund. Variations of these quality dimensions are used by all National Statistical Offices throughout the world. Two examples of this are: The Quality declaration of Statistics Netherland (2008) and Statistics Canada's Quality Assurance Framework (2002). There are 8 dimensions in SASQAF but we will be focusing a few of the dimensions that are directly linked to the benefits of digital data collection.

Quality Dimension	Digital Data Collection
<p>“The <i>relevance</i> of statistical information reflects the degree to which it meets the real needs of clients. It is concerned with whether the available information sheds light on the issues of most importance to users.”</p>	<p>Digital collection is flexible; however the relevance of information is informed by processes determining the user requirements. The linkage would be the flexibility of implementing the desired survey on the device. The determination of this dimension starts with defining user requirements before the survey planning and design starts. It then runs through to the tabulation and dissemination output.</p>
<p>“The <i>accuracy</i> of statistical information is the degree to which the output correctly describes the phenomena it was designed to measure. It relates to the closeness between the estimated and the true (unknown) values. Accuracy is measured by means of two major sources of error, namely, sampling error and non-sampling error.”</p>	<p>The use of digital data capturing will address some of the non-sampling errors that occur during a survey. Questions that are incorrectly answered on the device during capture phase can be addressed immediately, since the existence of wizards, helpful instructions and comments can coexist with “LIVE” capturing. Skip patterns are electronically enforced. The development and implementation of a detailed quality management plan will assist in this.</p>
<p>“The <i>timeliness</i> of statistical information refers to the delay between the reference points to which the information pertains, and the date on which the information becomes available. It also considers the frequency and punctuality of release. The timeliness of information will influence its relevance.”</p>	<p>This is where we perceive the biggest advantage of digital data collection would be for Statistical Organizations. Due to the design and implementation of the system, the data processing phase of the survey is reduced significantly. There will still be some elements of processing relating to cleaning, editing and imputation. The transfer of data directly after the interview to a data base saves a lot of time.</p> <p>This saving on time and processes means that the survey result can be released much closer to the end of the enumeration phase of the survey. The financial benefit encrypted in this is a huge advantage. However the fact that the data that is available is released closely after the reference point of the survey probably is of a bigger value to users. This however cannot be measured in a monetary value. User satisfaction might be able to provide</p>

	us with more insight into this fact.
“The <i>integrity</i> of statistical information refers to values and related practices that maintain users’ confidence in the agency producing statistics and ultimately in the statistical product.”	This dimension relates more to the management of the survey processes than the actual system itself. Training of fieldworkers and the transparency of the related processes plays a very important part in this. There is however some contribution that the actual system can make through added security measures that should enhance the confidence of both participants of the survey as well as users of the data.

5.3 Where is DDC heading?

The acceleration of the development of more powerful devices is going to open up new possibilities for DDC.

The devices are capable of shooting videos, playing videos, shooting high quality photographs, viewing a wide variety of photograph files, and recording and playback of voice. These multi-media functions could begin to reshape the techniques used for gathering information in the face to face interview process.

GPS capabilities of these devices mean that GPS tracking of enumerators is possible. It is possible to have beacons indicating on a map all the different enumerator’s locations. These beacons can display the number of interviews conducted by the enumerator, his contact details, his remaining battery life on the device etc. There are many possibilities.

GPS capability combined with a quality control system for checking the close to real time data flows will be powerful tools in locking down survey operations. Live monitoring of field workers and the data they are sending in will allow one to act swiftly to irregularities with regards to enumerator performance and data quality. Date/Time/Location stamps will prove to be useful in this regard. Furthermore, there will be flexibility in the assigning of surveys to different enumerators. A QA person will have the ability to send surveys to enumerators while they are in the field. Thus, if an enumerator falls ill, their work will be able to be reassigned to someone who is capable of working.

5.4 Some experimental design limitations:

The QLFS pilot was conducted using a small sample and did not have all the variables for measurement clearly defined. Thus, detailed investigation into the results was difficult. However, cumulative knowledge gained through previous exploratory tests combined with the QLFS pilot provided enough information to draw some interesting discussion.

5.5 Conclusion:

From a macro technological perspective it can be concluded that the time is right for DDC in large scale survey operations in developing countries. The question that remains is whether a system can be developed that will stand up to the rigors of international statistical standards and function effectively. This paper has sought to show that there are capable systems for DDC. However, the sizable benefits associated with the successful introduction of DDC in to large scale survey operations for households cannot be reaped without strong commitment from NSO's.

It is the conclusion of this paper that the time is right for DDC as it has worked successfully in the exploratory tests and gives a strong motivation for further testing of the system using larger samples and NSO collaboration.

References

- International Monetary Fund, (2003). *Data Quality Assessment Framework and Data Quality Program*. Available at: <http://www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm>
- Statistics Canada, (2002) *Statistics Canada's Quality Assurance Framework* Available at: <http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-586-X&CHROPG=1&lang=eng>
- Statistics Netherlands, (2008) *Quality declaration of Statistics Netherlands*. Available at: <http://www.cbs.nl/en-GB/menu/organisatie/kwaliteitsverklaring/default.htm>
- Statistics South Africa, (2008) *South African Statistical Quality Assessment Framework (SASQAF)*, National Statistical System Division, 4th draft, 2-3
- United Nations (UN), (2005) *Designing Household Survey Samples: Practical Guidelines*, Series F No.98, United Nations, New York