Secure Digital Data Collection In Household Surveys

Case Study: Statistics South Africa

Mahier Hattas (Author 1)
Department of Computing
University of South Africa (UNISA)
Pretoria, South Africa
mhattas@gmail.com

Prof. Mariki Eloff (Author 2)
Department of Computing
University of South Africa (UNISA)
Pretoria, South Africa
Eloffmm@unisa.ac.za

Abstract—**Digital Data Collection in South Africa is continuously evolving as technology and infrastructural networks gain momentum with respect to its development. In-field data collection is critical for any national government department who is mandated to supply the country and the international community with official data.**

The paper aims to illustrate the methods used by Statistics South Africa (StatsSA) in collecting household data using a digital collection process. The Quarterly Labour Force Survey (QLFS) and Dwelling Frame Project's are the primary focus areas of implementation. The paper further focuses on the background to the technology, its usage, problems encountered, lessons learnt, software/ questionnaire development, and more importantly the security issues around the collection of the data.

Keywords- Digital Data Collection; Household Surveys; Data Confidentiality and Security

I. Introduction

The use of digital data collection technologies is revolutionizing the way surveys are being done nationally and internationally. In keeping abreast with current methodologies and new technologies, StatsSA has decided to implement the use of digital technology for data collection in its household-based social surveys.

The long term goal is to eventually move towards a total digital system of data collection, thus totally eliminating the paper trail. This is in line with the organisation's vision of savings in terms of cost, turnaround time for data usage, confidentiality, security of data and the environmental factors. Since the digital system is re-usable and easily scalable, it is also viable as a sustainable solution. The paper addresses key ***security and quality assurance*** [1] issues of the organization. Confidential information is parsed from the field of collection into a secured central server at head office. In keeping up with current data collection methodologies and new technologies, StatsSA has implemented the use of GPS technology on handheld computers viz. Personal Digital Assistants (PDA's) in conducting; monitoring and reporting in its household surveys. The Geography division is the primary driving force behind this initiative which is quickly gaining momentum and stirring inputs and expectations amongst other divisions within the organization.

II. Background

The purpose of the paper is to prove that digital data collection for household-based surveys is feasible based on the QLFS and Dwelling frame experiences. Security of the data is key to ensuring data is not compromised or altered at any stage during the processing of data. Given the experience of the diverse challenges faced in the field; monitoring, tracking and management of surveys. The rapid stabilization in terms of support and infrastructure as well as the progression to pilot digital capturing of a survey questionnaire has all led to the realization that the time is nigh to start fully using the technological capabilities of the current systems within the organization for household-based surveys. The alternative plan B, viz. a paper solution is advised during the migration from any manual to digital process. [7]. Although solutions via various studies in different fields of applications e.g. clinical research, geographical geomapping, military geomapping, data acquisition, engineering, marketing etc. indicate that digital data collection is recommended broadly, however the value of information is a key factor in deciding the emphasis placed on the choice between manual collection vs. digital collection predominantly so if the organization is the official supplier of statistics to the country.

A. Quarterly Labour Force Survey (QLFS) and Dwelling Frame (DF)

QLFS - The Labour Force Survey is used to measure the health and wellness of the countries work force. It also measures employment and unemployment rates in the country [1]. Basically the survey consists of three abstract phase's viz. publicity, listing and data collection (also called enumeration). ***Publicity*** is the first step and it involves the process of communication to the communities/households that were sampled, informing them of the impending survey and the processes involved around it[1]. Secondly, ***listing*** is the process of circumnavigating the community and finding out what the area comprises of in terms of structures (private dwellings, shops, schools, institutions etc.). Other purposes include: collecting information on the number of people living in an area and what type of structures exist etc. [1]. Lastly, ***data collection*** (enumeration) is the actual interview phase where enumerators visit the selected sampled households [2].

Currently all survey fieldwork is conducted using paper questionnaires. During the pilot, fieldwork listing process, the QLFS enumerators managed to geo-reference approximately 86% of the entire paper based listing. As with all new technologies a few minor hiccups were encountered in implementing the system. These problems were the main reasons that the enumerators were not able to geo-reference a total listing (100%).

Problems encountered included:

- the lack of technical support for field related problems
- trying to reconcile paper and GPS records

DF - Having witnessed the potential of using digital data capture devices in the field, albeit for the purpose of geo-referencing only; the DF project decide to expand on the success of the QLFS pilot by utilizing additional elements of digital technology. The primary goal of the DF project is to geo-reference every dwelling in the country in order to facilitate the demarcation process for the 2011 Census. The DF project re-designed their methodologies in order to harness the full potential of the digital data collection. This process resulted in creation of a digital field data transfer system to facilitate the transfer of data captured by the enumerators using handheld PDA devices (via GPRS) directly into a secured central database located at StatsSA head office in Pretoria. In addition to this; a technical support call centre was set up to assist enumerators with any difficulties / troubleshooting / problems encountered whilst capturing in the field. In addition to the technical support call centre, digital data support specialists were appointed in each province with their primary function being the first-line field support for technical problems. At the time, the DF project has captured well over 2 million records with complete attributes. This information is sent instantaneously from the field to the central database in Pretoria via a GPRS connection. This eliminated the need for processing of geo-information since it was accessible instantaneously via a web based interface.

GPS

Global Position System or GPS (see Figure 1) , a global navigation satellite system was set up by the United States Department of Defence during the 1970s, for the purposes of warfare and has since then been adopted primarily for public use [3]. A GPS receiver calculates its position by measuring the distance between itself and three or more GPS satellites. Since the speed at which signal travels between satellites is known, by measuring the time delay between transmission and reception, the distance to each satellite can be calculated. The signals also carry information about the satellites location. By determining the position of, and distance to, at least three satellites, the receiver can compute its position using trilateration. Normal GPS accuracy is between 5 to 10 meters. Accuracy depends on a variety of factors such as whether you are in an open space or in a built up area as the signal can get distorted1.

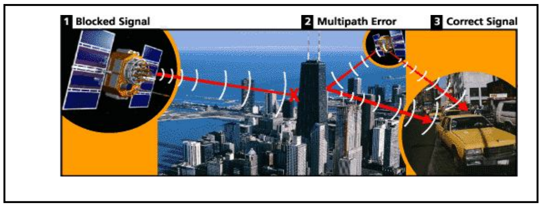


Figure 1: GPS signal reception

Device Selection

When Stats SA started the investigation in the feasibility of using the GPS and handheld devices, it discovered that there existed a variety of different devices offering different options with regards to common criteria e.g.: battery life span, memory, GPRS connection, etc. The investigation revealed that the organisation's GPS requirements would ultimately depend on the business needs of the organisation. Generically, all organizations would however need to take into account the following 3 main criteria:

- device specification e.g. rugged or standard PDA
- budget availability
- the required accuracy for the fieldwork application etc.

Since StatsSA collects household level data in a variety of geo-locations and social conditions, accuracy was the defining point. Accuracy was particularly important for informal dwellings. The location of a particular household allowed the creation of a geo-referenced database for households. As a result of this, it was decided that Stats SA would require an accuracy of 1m or less. Informal settlements are complex structures built virtually close to each other. The nature of these structures is such that it is mobile in appearance and any material is used to create a temporary structure. Using satellite imagery is not sufficient as many households tend to be occupying a single roof. This would result in an underestimate of households and household members and may create a possible bias during sampling selection. Therefore the 1 meter accurate GPS with additional attribute information is captured during the listing phase. Using Android-Tablets or other devices for digital data collection, although more cost effective, the accuracy (w.r.t Geo-Location) and additional capturing may not be effective in informal settlements, although it would be other areas.

A standard GPS machine can obtain an accuracy of around 3 meters on its own and a modified GPS needs to be purchased in order to increase the level of accuracy. Sub meter accuracy is usually obtained by means of a backpack which is connected to the GPS logger (see Figure 2) on one side and an antenna on the other. The logger is the actual GPS unit on which the information is captured whilst the antenna is used to communicate with the DGPS transmitters. This is unscrambled in the processor that is housed within the back pack and this then sends the sub meter accurate coordinates back to the logger which stores it on the data card or memory which can then be transmitted via GPRS or 3G using the existing telecommunications network or they can be stored on the device and downloaded onto a PC when back at the office. The data is encrypted before sent from the device to the server. Within the backpack are also batteries that power the whole system.

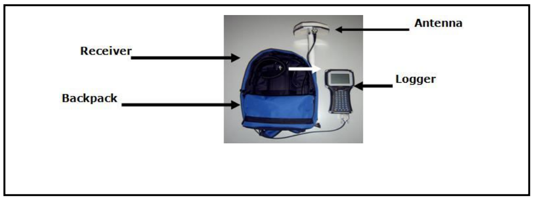


Figure 2: StatsSA's selected GPS Device

After thorough investigation of all the available options, the device chosen met the following criteria:

- Built in GPS receiver
- GPRS/EDGE connection to transmit information wirelessly
- Fast processing speed of 520mb (CPU, other PDA average 200mb)
- Memory (256MB, upgradeable with SD card)
- Alpha numeric keyboard (a bonus option to those that don't like the touch screen keyboard)
- Battery life of 8-10 hours

In addition to this, StatsSA bought the backpacks that enabled the devices to receive sub meter accurate signal. The organization also purchased the sub meter DGPS (Differential GPS) signal. Having the correct system setup was very challenging, however common usability [6] issues also impacts the collection of the data. In the ***initial stages,*** the complete system with the backup contained a combination of system and usability problems during enumeration and the following corrective measures and lessons learnt were taken into consideration i.e:

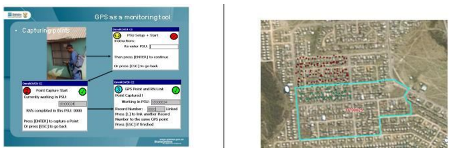
- *Data Cables and Connectors:* These were exchanged with the new ones from the office in the field.
- *Batteries:* Although sufficient batteries were received from StatsSA Head Office some were not tested before being sent to Provincial Office for use.
- *Signal or no fix error messages:* Currently there is no proper method of avoiding poor signal coverage areas. In such areas, enumerators were advised to use different times of the day, check for any obstructions and move around to obtain full signal. It must be noted that signal varies from area to area depending on the coverage and can also be affected by power supply i.e if batteries have insufficient power this can also lead to a no signal /fix errors.
- *GPRS configuration settings:* Have a tendency to switch off when the systems are not being used for longer period – Most of these were corrected in the field.
- *Charging the system:* Although this was covered during the training sessions, insufficient time was provided for device training. Two days were not sufficient for the GPS device training. Regarding chargers itself, they were made available in the province in the event faulty ones are sent back to the manufacturers for repairs.
- *Call Center / Helpdesk / Tech Support*: The initial call center could not effectively assist survey officers and there were no reference numbers to make follow up on who assisted in the troubleshooting of problems encountered during data collection. This was corrected as better tracking mechanisms were in place to ensure effective continuity of data collection and the functional use of tech support to enumerators at all times.

Overall, suggestions was that the training plan need to be changed and adaptable since the emphasis was more on the methodology, map reading and questionnaire rather than the practical application of the GPS system itself. More time was therefore allocated for the system training to the enumerators.

B. *How was the device used for QLFS?*

Enumerators used both handheld and paper based data collection methods. The devices were just used to capture GPS coordinate information. The associated attribute information was captured using the paper based listing form. When the enumerator first starts up the GPS (see Figure 3) they need to register on the device the PSU (Primary Sample Unit) number so there is a record of who was working with the device and in which area. This is useful in monitoring which enumerators are constantly experiencing problems and in which areas. The supervisor or area manager is then informed who can then investigate the problems/issues as they arise.

As soon as a enumerator reaches a structure, they are required to capture a GPS point. A GPS point is captured for every single structure, with the only exceptions being hostels or flats in which case a single GPS point is captured and the remaining units in the block of flats for example is linked to this single GPS points, viz. many records were linked to one point in this instance. Normally, only a single point is required because only one point will be needed to navigate back and collect data from all the people in the flat or hostel. The GPS point is usually captured as close to the front door as possible. The whole idea of capturing a GPS point is that the enumerators can come back later to the dwelling during the data collection phase and collect information.


GPS as a monitoring tool

- Capturing points

OmniROVER CE

1.2 PSU Setup + Start

Instructions:

Re-enter PSU:

Then press [ENTER] to continue

Or press [ESC] to go back

OmniROVER CE

2 Point Capture Start

Currently working in PSU:

10100024

RNs completed in this PSU: 0000

Press [ENTER] to capture a Point

Or press [ESC] to go back

OmniROVER CE

3 GPS Point and RN Link

Point Captured !

Working in PSU: 10100024

Record Number: 0001 Linked

Press [L] to link another Record Number to the same GPS point

Press [ESC] if finished

Figure 3: GPS Screen and PSU boundaries

Another major benefit of using a GPS is that for the first time Stats SA was able to actually physically monitor and see where enumerators were capturing information. By using the GPS points they were able to quickly identify areas in which the enumerators were capturing outside the boundaries (see Figure 4) or in the wrong area or even places where they did not capture all the information. In such cases, their supervisors or they were called and this was quickly rectified, thus saving time and increasing the validity and confidence in the data.

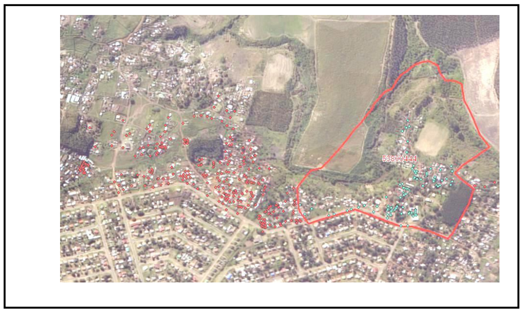


Figure 4: GPS Boundary and points

C. *Lessons learnt from QLFS*

As promising as this technology is, there are and always will be issues encountered when implementing new technologies. The first was the reluctance of users to accept that the technology would be useful, given the capital investment and other pilot studies done previously. Linking work done on paper to GPS points captured on devices was challenging. A system was designed on the principle that a pre-printed record number for each unit of information on paper would be the link to captured GPS point by a record number generated whenever a point was captured. The device generated the record number automatically thus the enumerator was tasked with checking every time that the number displayed on device was equal to the number of completed structures captured on paper. The system was not tested sufficiently and during listing it was found that enumerators were not diligently keeping track of paper trail against the device information.

D. *Digital capture pilot using Dwelling Frame digital data technology*

As mentioned earlier, the Dwelling Frame project built on success and lessons learnt from the QLFS project to incorporate a fully digital field data collection methodology. This included the development of the technical infrastructure to support wireless field data transfer.

The StatsSA: Western Cape provincial office recently tested the Dwelling Frame GPS system (NOMAD – See Figure 5) devices in capturing information for the QLFS survey as a pilot [4]. The QLFS questionnaire comprised of 60 pages containing information for approximately 6 individuals (respondents) per questionnaire. The results with respect to the digital interface; skip pattern effectiveness; data exporting into the pre-determined data template; enumerator adaptation to digital system; in-field response to device usage; speed of data transfer from the device to server etc. all contributed positively to the viability of digital data capture. The initial results were problematic due to the system specification being an adaptation of the Dwelling Frame system specifications.

E. *Questionnaire Design on PDA*

The graphical interface of a PDA digital device is compared to most of today's cell phone technologies. It couples user friendliness and robust usage. In comparison with paper enumeration, there is a definite limitation with respect to the size of the screen and the limitations of the questions designed on a device.

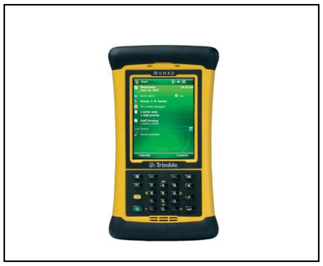


Figure 5: Nomad Device (R)

The main objective in the conversion from a paper form to digital form is not to lose the essence/logic of the question being asked such that potential bias in the results is minimized. Essentially, using a digital device to capture information should have the same response as if the data was collected using the paper trail. The QLFS questionnaire was satisfactorily converted onto the device without any major changes to the design of the questionnaire. The main problem though on the device was to cater for multiple respondents. The design of the questionnaire on the device was inadequate as the Dwelling Frame questionnaire on the device was adapted to suite the needs of the QLFS questionnaire for the pilot. These minor alterations was noted and reported by system developers to being a minor change that would be out of scope for this test. The automated skip patterns were found to work highly effectively as all possible paths through the survey logic were worked through by experienced DF enumerators. The answers received from the respondents, prompted from the digital interface, were consistent with those associated with the paper method.

F. *User friendliness and adaptation to the digital system*

The DF enumerators were trained on the devices over two sessions on separate days. Having the experience on DF devices, the enumerators adapted to the device and the interface quickly. After the training they independently navigated their way through the digital questionnaires with confidence. There were however a few enumerators who required more attention to the QLFS questionnaire than others on the devices. However, after the training sessions, all enumerators were proficient in the use of the device in conducting surveys. In terms of respondents, they were curious at first, but later became more comfortable as the enumerator conducted the interviews. In general, retraining on the device was very helpful for inexperience staff as things became much clearer during the practical use of the system. In areas where there was no/weak signal, the data collectors understood the processes of methodologically circumnavigating until a signal is acquired. If absolute no signal, then the device catered for offline input and information is securely sent via gprs to the server. The loss of GPS coordinates for the sampled structure is then a drawback on the system itself. Tech Support was present in the field assisting enumerators in troubleshooting and system-related issues. Quality Assurors and Supervisors also assisted in the methodology aspects with respect to data collection.

III. SECURITY AND DATA TRANSFER

In terms of security, according to SASQAF [1] "*The integrity of statistical information refers to values and related practices that maintain users " confidence in the agency producing statistics and ultimately in the statistical product*" data integrity is very important. Aspects of data security ranges from:

- ***Data Capture in field***
  - Human error can occur during capture-time, if the application is not well-designed to ensure logical skips and correct data entry, then the integrity of data may be compromised. The application on the device ensured that according to the business rules, all compulsory fields were captured and verified using lookup tables. Skipping patterns were followed and the device ensured completeness of data collection. Usernames and logins also ensured correct people expected to capture the information was maintained.
- ***Data Store on device***
  - If data is not encrypted on the device, it is open to virus attacks or scripts on the devices may risk it being lost or data being altered. The information was encrypted especially in the cases where there were no immediate synchronization with the server due to a lack of signal. However regarding the signal/No fix error messages, there is no proper way of avoiding poor signal coverage areas. In such areas the enumerator is advised to use different times of the day, check for any obstructions and move around to obtain full signal. It must be noted that signal varies from one place to another depending on the coverage and can also be affected by power supply i.e. if batteries have insufficient power this can also lead to a no signal / fix. The encrypted data on the device is temporarily held on the device and only transferred when in a gprs area.
- ***Data send to server***
  - Upon upload of data to the server, data was encrypted and sent in secured packets across the network. In transmission, in ensuring that no data was lost in transmission and received on the recipient side (client) without any error or distortion was critical and proved to be successful during the pilots.
- ***Data extracted from server***
  - Downloaded data from the server is another point of weakness if the server is not protected. Logins and authentication mechanisms were maintained for relevant uses with relevant levels of access to the data upon extracting, importing or downloading.
- ***Monitoring and quality assurance***
  - The data manager component allowed supervisors to track the progress of data collection real-time. Out of boundary issues and progress reporting allowed for easy monitoring of infield data collection. Setting up real-time monitoring in the field for quality control also aids security of devices in-field. Response to inactivity or downtime could be a result of data loss, theft or security breaches with respect to information transfer.
  - Software Management also allows the non-use of other functions on a PDA. Since security may be affected or the system may reduce in performance as a result of enumerators using applications not intended for use. The administrator settings will allow authorized override on these as well as the ability to client settings on any device to perform optimally or upload critical security fixes if need be. [9]
- ***Physical security***
  - The backpack used with the device was seen as a hindrance in some areas. Especially in the high-crime areas, the backpacks created some curiosity amongst the public. This could often lead to many unavoidable situations whereby the security of the person capturing of the data as well as the data itself is in jeopardy. Proper planning, publicity and the help of the police force ensured the physical during the field collection. Furthermore, the device is robust and contains protective covering.

The devices used in the pilots were not damaged; lost or stolen during enumeration. 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. In terms of data transfer, the device is synchronized with Head Office central operations and is designed to send the 'completed' questionnaires directly to a secured server as soon as they have been completed. Thus, as an enumerator finishes an interview it gets sent. The speed of sending the questionnaires to server averaged between 1 to 2 minutes. If there was no signal, data was stored and sent at the next time of possible synchronization.

A. Discussion

A greater number of processes are increasingly becoming digital, and with technology rapidly advancing, the evolution of a digital data collection process for surveys is the next logical step in keeping up with this advancement. The QLFS and Dwelling Frame projects have in many ways provided (and continues to provide) the testing ground for the use of new technologies that can be adapted for larger projects like a population census [5].

Security issues is a real concern, especially in the domain of official statistics. The development of ***security policies*** and frameworks being built to ease the minds around the validity of a paradigm shift to digital technologies is vital to the organisation. StatsSA have seen the value by using the basic concept of capturing points in order to monitor progress. However the technology is not limited to merely this and Stats SA is committed to ensuring a move towards a totally digital survey. This is however a long road that needs to be covered and we are just at the beginning of our journey. At the same time, ***backend development databases*** and programming of forms for capturing information are continuing. This will enable to enumerators to capture information on the device via drop down boxes or entering it using the alpha numeric keypad. The aim is to increase the amount of information that is captured on the device and utilise GPRS technology to facilitate data transfer directly from the field. There will be parallel process that will run with information being captured on paper as well as on the device but this would only be until such time that the development is completed and running smoothly, then the paper will be gradually phased out until it eventually disappears. The digital data collection process is no doubt a challenging and exciting project.

B. Conclusion

This paper has sought to show that there are perhaps capable systems and processes for digital collection at StatsSA. However; the sizable benefits associated with the successful introduction of digital data collection in large scale survey operations cannot be reaped without strong commitment from the organisation to move tactically and strategically towards technological systems. Security-related issues are real and information security is at the forefront of the drive to prove the validity and trustworthiness of a digital collection system with reinforced processes to ensure the integrity of its application. Heavy emphasis should be placed on device training and methodological training of completing any survey [9], and close monitoring of the performance of the system and all staff resources.

Other digital methods e.g. using web-based technology may not be useful in most of the South African areas. Lack of physical infrastructure and levels of access to internet is not as high as apposed to some developed countries. Key to the success of this technology is the linkages between a centralized IT infrastructure responsible for their core function of electronic data transmission, capture and security (to and from respondents) as well as the statistical expertise of the organizing statistical project/ programme heads. Integration and migration away from the manual paper collection is a goal realizable as is already benefited by economic clusters in the some developed countries like the USA [8]. Even in developed countries, access to internet is still optional to users and not enforced, hence web-based data collection lends itself to voluntary participation, which may be a problem for response rates to surveys.

Strong commitment also implies a paradigm shift in the way the organisation operates and thinks to embracing non traditional methodologies and operations. This coupled with up-to-date processes to ensure seamless integration towards digital surveys from paper-based collection; digital data collection could be utilized using a parallel transition methodology between manual (paper-based) and automatic (digital data collection) processes in the aim for an increase in automation and productivity. Other technologies such as smartphones have been tested previously within the Western Cape Statistics SA office. Some of the limitations were centered around technology build of the software solution, however as time progressed, the those technologies became more robust, secure and better ensured the confidentiality and integrity of data which is key for any statistical producing body.

References

[1] Statistics South Africa. (2008). ‘South African Statistical Quality Assessment Framework (SASQAF)’, National Statistical System Division, 4th draft, 2-3’

[2] Statistics South Africa. (2008). ‘Guide to the Quarterly Labour Force Survey’, http://www.statssa.gov.za/qlfs/docs/Quarterly_Labour_Force_Survey_Guide.pdf

[3] Wikipedia.(2009), http://en.wikipedia.org/wiki/Global_ Positioning_System

[4] Pilot test – (DF-QLFS Pilot) Digital Capture February 2009.

[5] U.S. CENSUS BUREAU. (2005). ‘Census and Survey Processing System’, http://www.census.gov/ipc/www/cspro/

[6] Olmsted, EL. 2004. ‘Usability Study on the Use of Handheld Devices to Collect Census Data.’ *IEEE* (2004)

[7] Guadagno, L., VandeWeerd, C., Stevens, D., Abraham, I., Paveza, GJ. & Fulmer. T. 2004. ‘Using PDA’s for Data Collection.’ *Applied Nursing Research,* vol. 17 (2004), pp. 283-291

[8] Swartz, R., Hancock, C. (2002) ‘Data collection through web-based technology’, Statistical Journal of the United Nations ECE, vol. 19, pp. 153-159.

[9] Shirima, K., Mukasa, O., Schellenberg, J., Manzi, F., John, D., Mushi, A., Mrisho, M., Tanner, M., Mshinda, H. & Schellenberg, D. 2007, , The use of personal digital assistants for data entry at the point of collection in a large household survey in southern Tanzania.