Open Data Kit: Implications for the Use of Smartphone Software Technology for Questionnaire Studies in International Development

By: Frances Jeffrey-Coker, Matt Basinger and Vijay Modi

During a study conducted in January 2010 by researchers of the Columbia University Mechanical Engineering Department in New York City, approximately 300 farmers were surveyed in rural Mali. Farmers were randomly sampled via standard proportional, stratified, cluster techniques. Data collection took place through the use of HTC G1 smartphones running Google's Android operating system. The phones were equipped with Open Data Kit (ODK) software; a system that immediately digitizes data for analysis, allows for remote monitoring of the collection progress, and facilitates the gathering of data, eliminating the need for paper surveys and therefore significantly reducing survey times. ODK has the potential for a profound impact on the future of data gathering, particularly in development applications where locations may be remote and budgets tight, yet where mobile phone use is rapidly increasing with the expansion of service coverage.

Introduction to the Technology

What is ODK?

ODK (Open Data Kit) Collect is an open source program in which programmed questionnaires are implemented on mobile smartphones. It is part of a suite of tools developed by a group at the University of Washington called Change, who explore the use of technology in improving lives in developing countries. The ODK questionnaires are written in xml format and can be created manually or generated automatically using an online interface. The ODK software suite consists of several different programs including ODK Collect and ODK Aggregate. The ODK Collect program is installed on a smartphone and the questionnaires are subsequently saved to the phone's SD memory, where it can be accessed and completed, even without wireless connectivity. ODK Collect enables users to ask questions not only

in a linear sequence but also with a predetermined if-then logic system, relying on answers to previous questions. The program also supports the incorporation of GPS points, photos, videos, bar codes, and sound bites as attachments to surveys or as the basis of the questionnaire responses. Then the surveys taken on the phone can be sent wirelessly to a server hosting ODK's Aggregate tool, via either wifi or a mobile internet connection. The other tools in the ODK suite, Manage, Validate, and Voice, were not used during this study.

Comparative Technologies

The use of mobile phones for rapid data transmission is not a new concept. Prior to conducting the study SMS based technologies were considered including RapidSMS and FrontlineSMS. These are also open platform tools that can post data to web-based servers via SMS messaging. However, these SMS based programs are more limited in their user interface flexibility and most practical with simple text information and short forms. RapidSMS, FrontlineSMS, and others can be used with any mobile phone without installation of special on-phone software, and for this reason they require low hardware overhead and are an ideal fit for studies that utilize short forms with fewer questions across many enumerators. Studies that require larger forms with more questions, fewer enumerators, and a variety of data types (i.e., GPS, photos, etc.) can benefit from more adaptive technologies, such as EpiSurveyor or Open Data Kit (ODK).

EpiSurveyor, ODK, and others utilize the features available on smartphones and allow for more complex question types. They can easily manage a large volume of information gathered in complex surveys, with less vulnerability to data entry error compared to SMS based technologies that can require tedious coding practices. Some of the more mature smartphone survey suites must be purchased on a per phone basis. Those platforms can be valuable for expansive, long term programs, requiring quick implementation. However, only free platforms such as EpiSurveyor and ODK were considered. EpiSurveyor was found to be a more mature platform than ODK, which was still in the process of being developed. However, the limited

variety of devices with which EpiSurveyor is compatible and the proprietary nature of these devices' operating systems led to the use ODK. ODK is compatible with smartphones running the Android operating system.

ODK forms easily support complex questions with a variety of data types including binary, integer, and string. The necessity of geospatial data gathering was also important for the study; ODK allowed for the acquisition of standard GPS points and their accuracies within questionnaires. For the Columbia University study, unlocked HTC G1 phones were selected.

Study Phases

Writing the Questionnaire

The questionnaire is created by writing an xml file that is saved directly to the phone. The survey is then implemented using ODK's Collect software. The files for this study were created using a free xml editor, but could also be edited using standard text editing programs on both Windows and Macintosh operating systems. Although the HTC G1 phone did not support French language menus, the survey itself was written in French for use in Mali. ODK Aggregate was hosted on appspot.com, the Google App Engine. The programmed survey was uploaded to the engine and all files transmitted from linked phones could be accessed from a computer connected to the web. Responses were downloaded from the the Google App Engine running ODK Aggregate in the form of delimited .csv files.

Field Implementation

Upon arrival in Mali, the phones were configured for access to mobile internet through a local telephone service provider. Configuration was a simple process of buying SIM cards and changing network settings. Thereafter, credit for data transfers was purchased with prepaid phone cards.

During Columbia's study, farmers were asked more than 100 questions concerning household demographics, personal histories, and farming practices, and GPS points were taken at their home and fields. One questionnaire took 20-30 minutes to complete. The farmers were interviewed by enumerators who worked for a local Malian Company – Mali Biocarburant. Mali Biocarburant (MBSA) is a biodiesel company that employs a network of local agriculture extension agents to work directly with farmers growing the biofuel feedstock purchased by MBSA for the production of biodiesel. The enumerators used in this study were local MBSA agriculture extension agents. These enumerators directly and independently operated the G1 phones and translated the questions from French to Bambara. They were trained how to use ODK Collect on the phones and how to download new surveys from ODK Aggregate while in the field. During the training phase of the

study, while computers were available, the enumerators also reinstalled software and download edited surveys via USB.

Given the widespread familiarity with mobile phones in developing countries, the enumerators were quickly able to navigate the G1 touch screens and within only a few trials they had cut in half the time required to conduct interviews. During the training and monitoring phase, frequent telephone correspondence with researchers in New York confirmed that data was being received on the server. This communication allowed for instant feedback to the field from the states, permitting immediate notification of progress and irregularities.

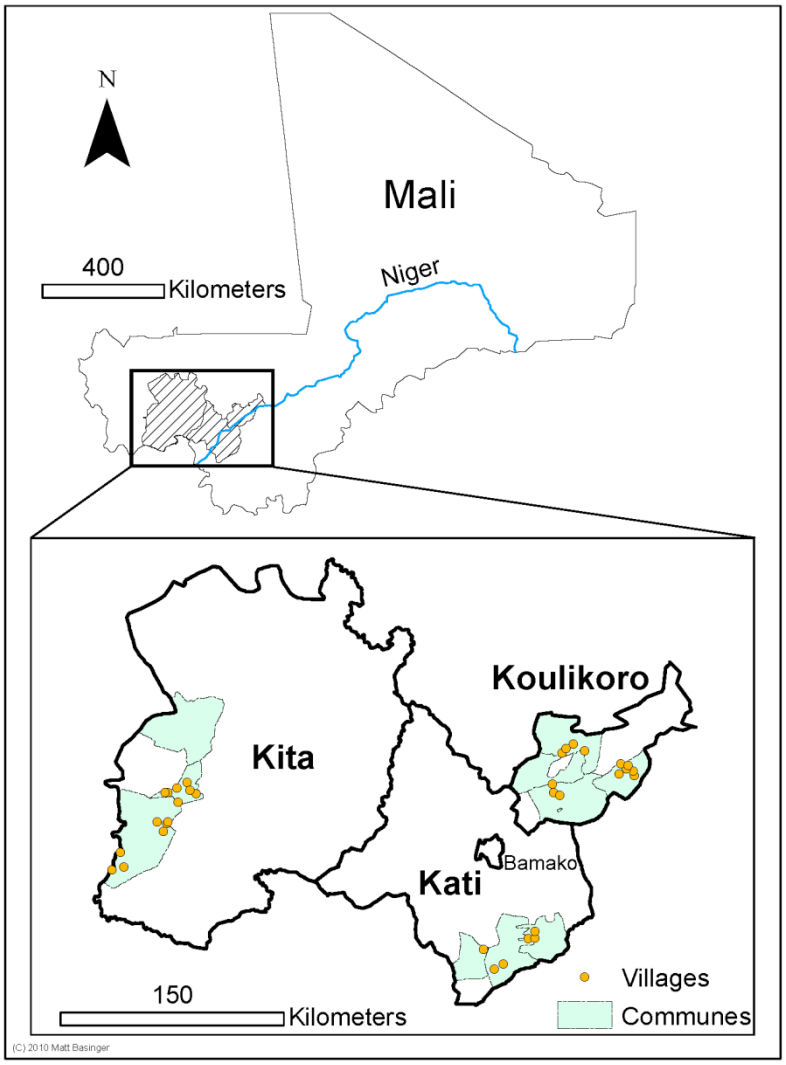

Figure 1: Map of Area Where Data Collection Occured

Remote Monitoring

The bulk of the survey-taking took place by the enumerators after the Columbia University researcher had left the country. As data continued to be collected independently, researchers were able to monitor real-time progress over the internet. Via ODK aggregate, unique phone IDs were associated with each posted

response. These IDs were used to differentiate between data sent from three different phones utilized in the three regions where data was collected. Figure 1 shows a map depicting the three regions where the survey took place. The newly uploaded survey responses were monitored by regularly downloading csv files generated by ODK Aggregate. Daily progress was observed and the enumerators contacted in order to comment on the data received. The enumerators also frequently contacted New York to request clarification or to ensure that data had been successfully sent. Once survey-taking was completed, immediate data analysis began with no additional digitization or organization necessary.

Retrospective Discussion

Limitations

Various limitations of the technology were encountered during the Mali study. Maintaining battery power in rural areas is an important factor to consider, as many large-screen smartphones have short battery life. In anticipation of this problem, additional high-capacity batteries and chargers were given to the enumerators. Battery capacity proved to be much less of a problem than feared because of local familiarity with battery charging solutions, due to the enumerators' own frequent use of personal mobile phones. However, the additional batteries did provide backup during occasional emergencies.

Local telecommunications infrastructure also exceeded expectations and in most cases mobile phone service was readily accessible, if not in every village then in a nearby area. Due to widespread use of phones, people living in the villages knew of exact spots where service could be accessed. However, in many instances the signal would not be strong enough to transmit survey data, and on a few occasions data was received more than once or not at all because of interruptions. As a result, constant communication was required between New York and Mali to verify data transfers. For projects abroad, internet access capabilities are not guaranteed, and verification of local service options is important before arriving in-country.

Furthermore, complimentary ODK applications, such as *Kobo Postprocessor*, should be considered by researchers looking to use ODK where no mobile phone internet service is available. *Kobo Postprocessor* allows for offline data-syncing: phones connected to a laptop can directly upload completed questionnaires to ODK Aggregate on the computer.

Overall, the limitations expected and encountered during the study were easily managed and proved less problematic than anticipated. No hardware problems were encountered despite rugged use, although more risks would have existed during the rainy season or the dusty end of the summer. Preventatively, an extra phone was kept on hand. ODK can keep multiple surveys saved to the phone, so in the absence of a network enumerators continued with surveys for a few days and then returned to a signal-accessible area to send several days worth of completed questionnaires all at once.

Problems Encountered

Various software problems were encountered during the study implementation process, though it is unclear whether or not they should be attributed to the Android operating system, the ODK software, coding mistakes in the XML questionnaire forms generated by the Columbia University researchers, or a combination of the three. Researchers looking to use early versions of ODK should be mindful of ways to work around these potential pitfalls.

One enumerator working in a remote area did not have regular access to either a phone network or a power outlet, so while accumulating a saved list of surveys, a back-up battery was required to continue with work before finding a location to send the data. However, each time that the enumerator replaced the battery and restarted the phone, all the saved surveys were lost. This problem was solved by noticing that for the first several seconds after the phones were started up, the SD card was not yet loaded. If the ODK software was launched before the SD card had finished loading, any unsent surveys were erased. To overcome this issue, enumerators waited a few minutes after starting the phone before launching ODK.

Mobile phone service expenditures significantly exceeded the expected costs because the HTC G1 phones would constantly attempt to access the internet, even with obvious connection settings turned off. It was not clear how to manage the settings such that the phones would not routinely attempt to connect of their own accord. Furthermore, additional phone credit was spent when internet programs were accidentally initiated. This problem was complicated by the fact that the enumerators could not read the English menus. In areas with less network availability this was not a significant problem, but in more urban areas, buying phone credit became the most significant expenditure, surpassing even fuel costs. In Mali, at the time of the study, the approximately 100Kb file size for each survey response cost about 25 cents to send, but in areas with good reception four dollars could easily be lost in a day through inadvertent connections..

Purchasing a phone plan that provides data transfer limits over a monthly period is a potentially more economical solution in face of the connection problem, but such service is not necessarily possible with certain providers and was not available in Mali at the time of this study. Therefore, different strategies can be employed to limit access. Phones can be turned off at all times accept when collecting or transmitting questionnaire responses. Additionally, small increments of phone credit can be purchased and added to the phone at the end of the day, once a list of surveys has been taken. Then they can be sent collectively using just the right amount of credit required. Removing internet-accessing program icons from the home screen also lessens the likelihood of accidental program connection.

A couple other connectivity issues were encountered during the study. GPS access with the phones was very slow in new areas, with latitude and longitude calculation times of up to ten minutes in order to reach an acceptable accuracy. Therefore with frequent and varied relocations, the GPS process could be time-consuming. Additionally the ODK software would simply freeze after a prolonged period of signal searching. The pop-up GPS initiation window had to be closed, the survey reopened, and the point retaken in instances of this timeout.

The phones also experienced occasional malfunctions when sending data in unstable network areas. Many duplicates of files were received on the appspot server running ODK Aggregate. These duplicate responses were assumed to be due to a service interruption preventing the software from verifying the already successful upload of the information transferred to the server, resulting in a repeated attempt to send the data until verification occurred. Likewise, sometimes files were not sent when the phone indicated that they had been. This occurred rarely and the missing surveys were found stored on the SD card for computerized sending, but it required that the enumerator email raw data (XML forms) directly. Fortunately ODK Aggregate allows the direct uploading of XML files from a PC into the server database. So in these instances, responses could still be easily and quickly integrated into the overall data set.

Lastly, the use of accented letters and special symbols in the French version of the questionnaire proved problematic. The xml survey files had to be saved using Unicode format with the appropriate file header intake because the ASCI special character definitions normally used by text editing programs were not supported on the phones. Standard Windows and Macintosh text editors both support saving simple text files in Unicode with a change of settings, but unfortunately even the Unicode surveys, once uploaded to ODK Aggregate, did not properly display symbols if the survey had been downloaded to the phone from the internet. If Unicode forms were uploaded directly from a PC they did not encounter this problem. In one instance a survey form was accidentally deleted while in the field, and the enumerator used ODK Collect to re-download it online. However, it was still missing all special characters. The issue was resolved by accessing a computer in the city to transfer the new survey via USB. (ODK developers later noted that another solution to this problem is URL encoding the forms using the specific HTML numbers corresponding to the special characters desired.)

Conclusion

Both domestic and international projects can benefit from the use of ODK as a data collection tool for household questionnaire surveys. Provided the financial capacity to buy the expensive smartphones and data plan, the ODK platform enables users to capture and instantly digitize information of a variety of formats, eliminating the need for paper questionnaire surveys. A certain level of technical expertise is required for writing the XML forms used to to create the questionnaires. Additionally, setting up a server space for aggregating responses also takes significant technical capacity. Once forms are written and a server established the subsequent use of ODK is very straightforward. Because the platform enables real-time remote monitoring of the data collection process, it is ideal for studies in developing or remote areas. If enumerators have questions about their data, the researcher can provide feedback from the information they have sent online.

Because ODK was created relatively recently, it is important to allow ample time to familiarize oneself with the software and to deal with technical problems before starting a project. At the time of this study, ODK Aggregate had not yet reached

version 1. Even though the ODK development team is very responsive to questions posted on their online forums (responses were often supplied in a matter of minutes or hours), more extensive documentation is needed for "non-developers" (i.e., non software programmers) to be able to seamlessly use ODK in the same way that a "non-developer" can use a more mature platform like EpiSurveyor. At the time of this study, the survey questionnaires still needed to be written manually in XML, although ODK developers are currently devising a platform for automated GUI-based generation of XML code.

In the meantime, with a reasonable level of technical expertise, researchers will find ODK to be a convenient and helpful tool in the implementation of questionnaire-based studies. The software has the potential to make a significant impact on the process of survey-taking for development applications due to its exceptional versatility and free open-platform nature, providing an innovative means for gathering information anywhere in the world and compiling it online. Because ODK is tied to the Android operating system, if Android based phones continue to grow in popularity and decrease in price, ODK may even begin to displace SMS based platforms that have claimed hold of project spaces where hardware prices and ubiquitousness are limiting factors.

Acknowledgments:

The authors thank Columbia University's Vijay Modi, Matt Berg, and Jiehua Chen for their oversight throughout the study. The authors also warmly thank Mali Biocarburant staff and administrators for their collaboration, especially Koreissi Toure, F.S. Rodriguez-Sanchez, and Timothy Singer. Special thanks is extended to the University of Washington researchers responsible for creating and supporting ODK, especially Yaw Anokwa and Waylon Brunette. The data collected via the ODK platform for the described study is currently being analyzed and written up for submission into a peer reviewed, scientific journal publication.