Open Source Data Collection in the Developing World

Yaw Anokwa, Carl Hartung, Waylon Brunette, and Gaetano Borriello, *University of Washington*

Adam Lerer, *Massachusetts Institute of Technology*

Open Data Kit enables timely and efficient data collection on cell phones, a much-needed service in the developing world.

In the developed world, data is relatively easy to collect. Be it population demographics, embedded traffic sensors, or even popular Internet services, the ability to easily tap and synthesize raw data enables individuals and organizations to make decisions. Examples of such synthesis include earthquake sensing via Twitter, traffic mapping in Google Maps, and disease-oriented websites like PatientsLikeMe and Google Flu Trends.

In the developing world, the lack of reliable infrastructure, ubiquitous connectivity, and adequate expertise makes data collection difficult. Currently, most organizations collect data on paper forms despite inefficiencies such as the physical collection of completed forms, data transcription errors, and long delays before the data is available.

This problem is exacerbated by the data's critical nature. If, for example, you don't know how far villagers are from a stagnant water source, it's difficult to know how many mosquito nets to deploy; and, if deployment information isn't connected to malaria cases at local clinics, it's impossible to know whether the nets have made a difference.

The exponential growth of cell phone usage and infrastructure in developing regions has aroused great excitement for using mobile devices to address current gaps in data gathering. In addition to the variety of data—text, photos, location, audio, video, barcode scans—that can be gathered, mobile devices have proven to be dramatically faster at both collecting the data and making it available to decision makers. Moreover, deploying mobile devices can be less expensive and less error prone than using pen and paper.

DATA COLLECTION

While several systems currently exist for simple data collection in developing regions, they're often difficult to deploy, hard to use, complicated to scale, and rarely customizable or extensible.

Current offerings like Pendragon Forms, Frontline Forms, and Nokia Data Gathering are inflexible because they're closed source and based on closed standards. Others like JavaRosa, RapidSMS, FrontlineSMS, and EpiHandy are more flexible but primarily collect textual data.

Moreover, many of the devices that run this software have limited processing power and restricted storage, and they often lack cellular connectivity. Input often comes in the form of a stylus or numerical keypad, while output must fit on minuscule screens—a combination that results in poor usability. In addition, developers only have limited access to the phone's resources, making it difficult to include essential inputs like the phone's unique identifier, GPS location, or captured photos with the data.

There's also a need to develop more and better server-side tools. Ideally, these tools should be as service-oriented as e-mail has become. In the same way that consumers no longer need to configure and maintain mail servers, organizations that collect data need "e-mail easy" solutions that let them ignore the hidden costs of server infrastructure: power, connectivity, maintenance, security. And just like e-mail, it must be easy to move the data across various systems.

OPEN DATA KIT

To help fill this gap, we are developing Open Data Kit (http://code.google.com/p/open-data-kit), a suite of tools that enables users to collect their own rich data. ODK is designed to let users own, visualize, and share data without the difficulties of setting up and maintaining servers. The tools are easy to use, deploy, and scale. They also go beyond open source—they're based on open standards and supported by a larger community.

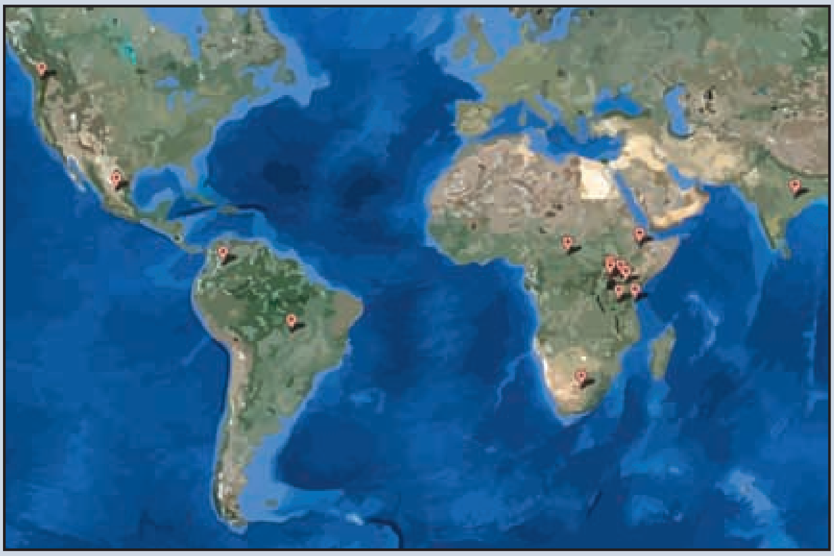

Figure 1. Examples of ODK deployments include deforestation monitoring in the Amazon, decision support for pediatrics patients in Tanzania, documenting war crimes in the Central African Republic, and monitoring school attendance in India.

ODK's goals are threefold:

- make tools modular and customizable so that they can be easily composed into appropriate arrangements for each deployment;
- exploit open interfaces and standards so that solutions are not "siloed" into monolithic enterprise-level packages that are difficult to understand and maintain; and
- establish data collection tools at the cutting edge of technology so as to avoid early obsolescence and make it easier to attract talented developers

Thanks to support from Google, we've already released a minimum set of tools required for practitioners in the field to begin collecting rich data sets.

ODK Collect, our client on Google's open source Android platform, renders a form, survey, or algorithm into a sequence of input prompts that provide navigation logic, entry constraints, and repeating substructures. Forms are based on the World Wide Web Consortium's XForms standard (and the OpenRosa Consortium's subset) and support a wide variety of data types, including GPS coordinates, photos, audio, video, and barcodes. The data entered is stored on the phone for asynchronous transfer via the General Packet Radio Service (GPRS), Wi-Fi, or USB cable to any XForms-compatible server.

ODK Aggregate is a ready-to-deploy server that hosts forms and submitted results. It aggregates collected data and provides standard interfaces to extract data such as spreadsheets, queries, and maps, and integrates with other systems via real-time Web requests. ODK Aggregate is currently implemented on Google's App Engine, enabling users to avoid the challenges of building their own reliable and scalable Web service.

ODK Manage allows the remote management of multiple phones. It ensures that appropriate forms and data files and applications are downloaded to each device without requiring user intervention. It also presents a dashboard to a supervisor for quickly disseminating updates and browsing the status of each phone in a deployment.

Although these three tools have only been available for download for a few months, uptake has been fast and broad, as shown in Figure 1. Examples include deforestation monitoring in the Amazon, decision support for pediatrics patients in Tanzania, documenting war crimes in the Central African Republic, and monitoring school attendance in India.

In all of these cases, ODK has enabled workers to be more efficient in gathering actionable data. However, the deployment that best demonstrates the power of the platform is an HIV treatment program in western Kenya.

AMPATH KENYA

The Academic Model for the Prevention and Treatment of HIV (AMPATH) is the largest HIV treatment program in sub-Saharan Africa and is Kenya's most comprehensive initiative to combat the virus. In mid-2009, AMPATH began scaling up a Home-Based Counseling and Testing (HCT) program to drive down infection rates. The idea behind HCT is simple—to prevent the spread of AIDS, AMPATH counselors must survey, counsel, screen, and test every one of the two million people in their catchment area.

Counselors currently attempt to visit 8-10 households a day, some with as many as 15 people each to counsel and test. They carry a Palm Pilot, a phone, a GPS unit, HIV tests, medical supplies, and many sheets of paper, and often must walk many miles between dwellings.

Upon reaching a household, counselors record its GPS coordinates and each inhabitant's demographic information and begin the counseling, testing, and screening session. They assign every household member a barcoded ID card to track any medical care provided by local clinics.

In cases where the HIV tests or the malaria and tuberculosis screening tests are positive, counselors refer those persons to a local clinic. For those unable to travel to the clinic, they also schedule follow-ups. Counselors manually synchronize all collected data to a Microsoft Office Access database and then transfer it to the OpenMRS medical record system. They also track all follow-ups and referrals on paper.

After considering several existing products, AMPATH has opted to use ODK to scale the HCT program's technology for several reasons:

- Surveys can be downloaded from and sent to OpenMRS directly over GPRS, avoiding the extra step of going through the Access database. Counselors need not wait for days to get data from the field to the hospital.
- GPS coordinates take seconds instead of minutes and no longer require a separate GPS device from which numbers must be read and reentered.
- Counselors can scan barcoded ID cards into their phone in a few seconds, as shown in Figure 2, instead of having to type in the data, which is time-consuming and can result in potentially serious data entry errors.
- An all-in-one device minimizes the amount of equipment, batteries, and cables counselors must carry from house to house.
- Referrals and follow-ups can be automatically managed. Counselors can send reminders directly to patients and place follow-ups on their phone calendars.
- Forms can include video, audio, and images to assist counseling sessions.

Over the next two years, HCT counselors will use 300 Android phones equipped with ODK Collect to reach millions of people in Western Kenya. They will be free to customize and extend ODK as they see fit, composing the tools into an overall process that supports their work.

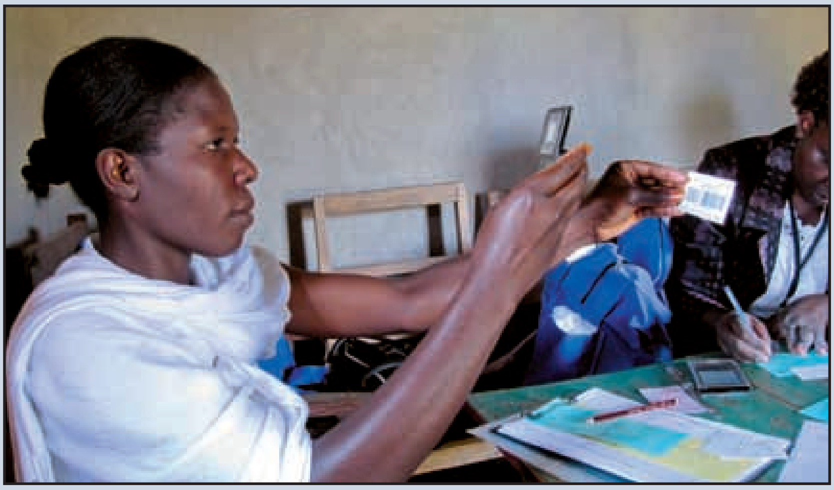

Figure 2. An HCT counselor scans a patient's demographic information into ODK Collect.

The ability to collect data is key to the success of many organizations operating in the developing world. Given the weaknesses of current tools and the surge in mobile phone growth, there's an opportunity for mobile and cloud technologies to enable timely and efficient data collection and thus change how healthcare is delivered to millions of people.

While ODK's current tools—Collect, Aggregate, and Manage—are sufficient for many organizations, we continue to build tools that push beyond what can be done with data collection today. These include voice-based forms, automated workflow management, and offline database replication. Although in preliminary stages, users who have seen the new tools are excited about the future of open source data collection in the developing world. ◼

***Yaw Anokwa** is a PhD student in the Department of Computer Science and Engineering at the University of Washington. Contact him at yanokwa@cse.washington.edu.*

***Carl Hartung** is a PhD student in the Department of Computer Science and Engineering at the University of Washington. Contact him at chartung@cse.washington.edu.*

***Waylon Brunette** is a PhD student in the Department of Computer Science and Engineering at the University of Washington. Contact him at wrb@cse.washington.edu.*

***Gaetano Borriello** is the Jerre D. Noe Professor of Computer Science and Engineering at the University of Washington. Contact him at gaetano@cse.washington.edu.*

***Adam Lerer** is an MEng student in the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology. Contact him at alerer@mit.edu.*

Editor: Bill N. Schilit, Google; schilit@computer.org