Leveraging technology to streamline the collection and visualization of agricultural and socio-economic data of vulnerable communities.

Juan Ardila, Juan Rojas, Felipe Ramírez, Rafael Vargas
United Nations Office on Drugs and Crime, Monitoring of Alternative Development
Cll 102 #17A-61
Bogota, Colombia
juan.ardila@unodc.org

DOI: 10.1481/icasVII.2016.f39d

ABSTRACT

This paper presents the use modern technologies implemented by UNODC in the monitoring of Alternative Development projects established in social and economically vulnerable agricultural areas in Colombia. The system supports data capture applications implemented via smartphones with the ODK and GeoODK arquitecture. Collected data include georeferenced field boundaries, field areas, geotagged photos, and socio-economic questionnaires. Offline collected data is stored in a relational database which is directly queried by several web applications that publishes and visualizes data reports and indicators in almost real time. A key feature of the designed infrastructure is the tight integration between data collection, processing, storing and visualization that ensures straight access to the raw data and derived indicators. The system is a low-tech solution and has been automated in order to secure data collection and minimize data loss or input errors. We present the system as an example of efficient information systems implemented by institutions working in developing countries with limited infrastructure.

Keywords: data collection, survey, monitoring, information, ODK.

1. Introduction

Currently more than 3.2 million people have access to the Internet and almost half of the world population has access to mobile broadband subscriptions (International Telecommunication Union, 2015). A large share of the adoption of communication technologies is happening in developing and least developed nations, which provide thriving conditions in places where physical infrastructure still limits the exchange of goods and until recently, ideas. As shown in Figure 1, as of 2015, the percentage of mobile broadband subscriptions reached 86.7% in developed countries and about 40% in developing countries (Dutta, 2009). There is, however, an increasing technology divide between developing countries and the rest of the world as shown also in the figure (left) (International Telecommunication Union, 2015).

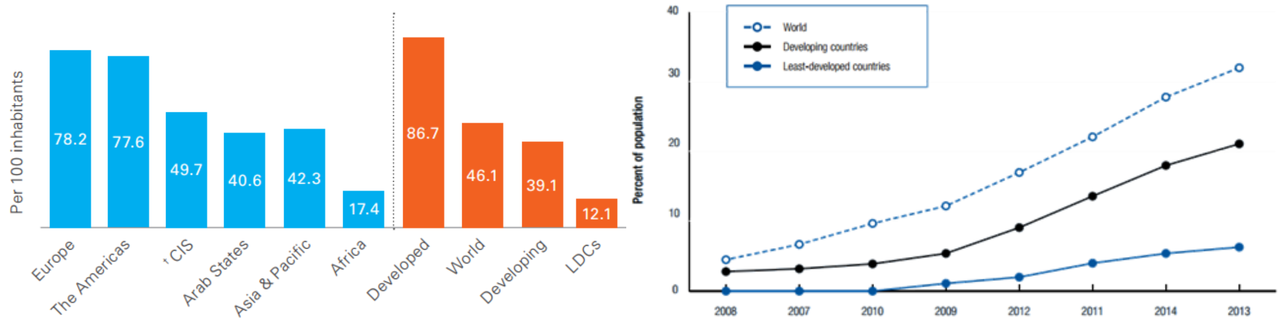

Figure 1. Left: percentage of mobile broadband subscriptions in 2015. Right: Mobile broadband penetration over the last years by region

As expressed by ITU, societies that transitioned to information technologies are experiencing a positive impact on economics, government, businesses and education, moving faster to the achievement of United Nations Sustainable Development Goals. The steady growing penetration rates of mobile broadband subscriptions in developing countries over the last years indicates that there are excellent business opportunities for the private sector but also enabling conditions for government and institutions to improve the knowledge of territorial and socioeconomic indicators of its populations.

Developing nations still face the legacy of paper loaded processes, red tape and slow data collection rates, which ultimately compromise the possibilities of data analysis and data dissemination to the final users. In this environment, data collection is so difficult, that often, completed questionnaires are dumped into boxes awaiting the availability of human and financial resources to interpret and make sense of their information. Publication of collected data often comes late, when decisions have been made, attention has already shifted to another project, and funds are depleted.

Recent ICAS conferences have reported efforts on the use of ICT systems to collect data and implement agricultural censuses. Those undertakings reflect specially the state of the art in developed countries or in national statistical agencies that have the infrastructure to invest on these endeavours. However, as exemplified by this paper, the development of infrastructure, communication networks, user communities, global positioning systems, software and hardware facilitates today reliable data collection, processing and dissemination at a low cost and a fast rate.

This paper reflects on the implementation of the information cycle of data collection, data processing and data visualization for the collection of agricultural and socio economic data. These processes have been implemented by UNODC in Colombia to document rural development transformations in areas affected by illicit crops. The experience is implemented in vulnerable remote rural areas, which largely match the conditions of least developing countries. As presented in this report, the developed system has resulted in timely data collection in areas monitored by UNODC in Colombia and an advanced organizational know-how of information systems.

2. The information cycle

UNODC supports the Colombian Government in the monitoring of the implementation of rural development programmes aimed at ensuring food security and socio economic integration of vulnerable communities immersed in areas affected by illicit crop cultivation (**insert note**). UNODC has been providing this support for more than a decade under several government administrations and policy implementations. The solution presented in this paper is the result of the evolution of a process that started in the early 2000's and has been improved and adjusted to honour the conditions of the policy implementation and the situation on the ground.

Under the Alternative Development monitoring programme, UNODC verifies: 1) the physical transformation of the territories being supported by the Government, and 2) the socio economic transformation of the communities targeted by the Alternative Development programme. This requires the collection of the following data in the field:

1) Measurement and georreferencing of parcels supported by the Alternative Development Programme. This includes the registration of the parcel boundaries, and capture of photographs of the area. Enumerators also collect data on land characteristics such as land ownership, land suitability and planted crops.

2) Administration of interviewing questionnaires inquiring about social, economic, environmental and institutional indicators related to the rural district of the participating household.

In the interventions of 2012, 2013 and 2014 a total of 34,000 households were registered in the Programme. Each household parcel was measured while questionnaires were administered randomly to selected households based on a stratified sample. Parcels were measured and identified at the beginning of the intervention, while interviewing questionnaires were administered during field visits at the beginning, mid-term and final stage of the project. This temporality permitted to monitor territorial performance of the communities involved in the program.

The information cycle of the implemented agricultural survey includes all the necesary steps to ensure: a) efficient data collection, b) data processing, c) analisis of derived information and d) publication and visualization of results. The transition from one stage to the next is an interrelated process, tightly integrated with coupled by technology, software, users and infrastructure (Figure 2).

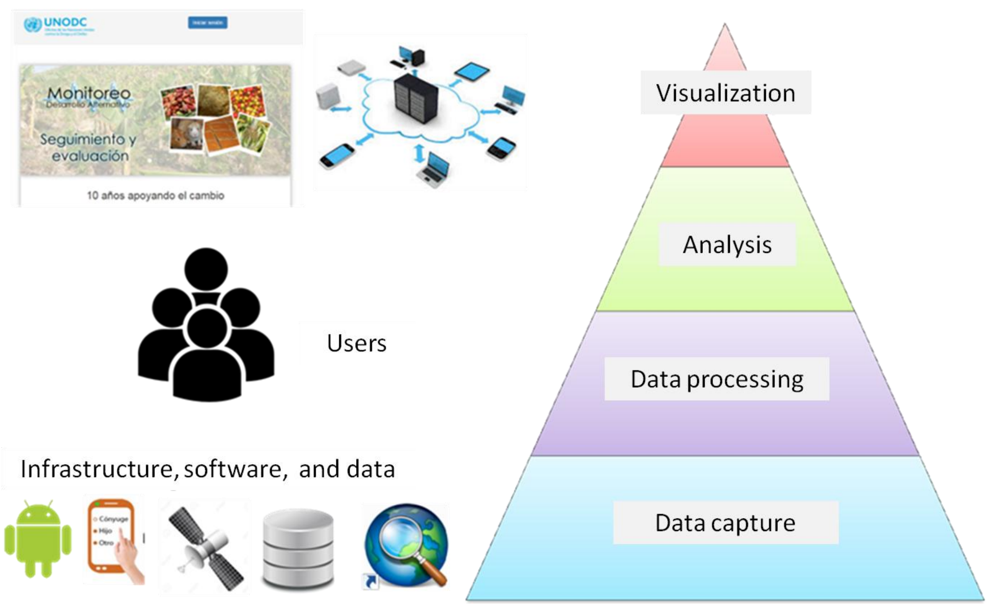

Figure 2. Stages of the information cycle: from data collection to visualization and dissemination

3. Data collection

We based data collection on the Open Data Kit project developed at the University of Washington (Hartung et al., 2010). This project includes several applications that assist during data collection:

ODK design: language and interface to design survey questionnaires. Questionnaires are implemented in the XLM/Xform standard which allows interoperability and readability by various systems. To facilitate multiuser form design, we used Google spreadsheets. Questions and logical flow of the questionnaire were validated and translated into an XLM file using the XLSform conversion tool.

ODK collect: ODK collect is the client application deployed to Android systems (currently supporting Android versions > 4.1). This application provides the ability to log into a specified server to download blank forms using defined user credentials. The server hosts the blank forms as well as completed submissions sent by collectors as explained below.

With the blank forms downloaded to the Android device, the enumerator can initiate data collection. Several forms can be stored in the same device, and with forms having a size of a few kilobytes, there is no practical restriction to the volume of collected data that can be stored in a device. The application interprets the questions, the flow and the logic conditions implemented in the form. For instance, the application prevents enumerators from leaving unanswered mandatory questions, or moving to the next question if some logical rule is violated with the current answer. Similarly, the application skips irrelevant questions depending on the answers of the interviewed household.

Collection of Geographic data: to enable spatial data collection we used GeoODK, a modified version of ODK Collect (Nordling, 2015). GeoODK facilitates the collection of polygons using the Geotrace geometry, where polygon vertices are recorded in continuous mode every few seconds or, alternatively, when requested by the user. Point data is also collected to register the location of the interview.

Location accuracy depends on the GPS chipset of the recording device being used. With the implemented system, any Smartphone device equipped with a built-in GPS can be used to collect the alphanumeric and spatial data. To streamline technical support, two devices were selected, based on their robust construction, GPS reliability, battery duration and cost (Figure 3).

Figure 3. View of the Monterra Garming GPS and Moto G Android smarthphone used for data collection

4. Data processing

Storage of collected spatial and alphanumeric data was centralized via the ODK Aggregate application. ODK Aggregate provides a ready-to-deploy server and data repository to: 1) retrieve blank questionnaire forms to be used in ODK Collect; 2) Accept finalized forms (submissions) from ODK Collect and manage collected data; 3) Visualize the collected data using maps and simple graphs; 4) export data (e.g., to a CSV files for spreadsheets, SQL database or as KML files for Google Earth).

ODK Aggregate can be deployed in several configurations, such as Google's App Engine, Amazon web servers or the organization server. Given the ease of configuration and server reliability needed for intensive and continuous data collection, we set up a hybrid system that stores collected data by mobile devices into an SQL database hosted by Google's App Engine and continuously synchronizes to a local repository.

We developed an R routine to process the data, normalize it and reshape it to the required format and structure of the local SQL database system. The script reads the CSV data, checks for new information and reshapes the data accordingly. The R system communicates directly with the database management system and stores the new records into the relational tables of the UNODC database (Figure 4).

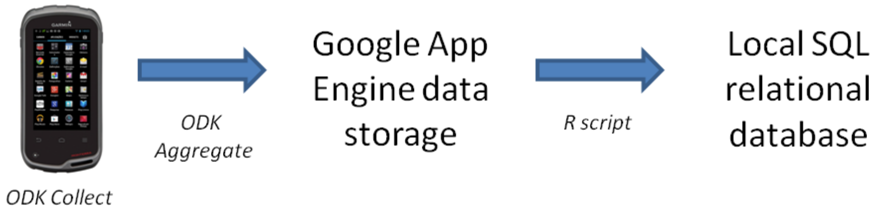

Figure 4. Flow of data from collection to storage into an SQL database

The spatial data collected in ODK undergoes the same transformation as the rest of the data described previously, but required a different treatment during the transformation in the R script. Spatial data is registered in the XLM form as an alphanumeric data type noted by a sequence of vertex coordinates in WGS84 system. The R script implements the following operations:

- Verifies and fixes topological errors: this involves removing signal overshoots and spiky vertices affected by the quality of the GPS signal during data capture.
- Verifies and closes open rings to transform them into polygons.
- Verifies spatial data consistency. i.e polygons should be located in a given municipality.

5. Data visualization

The automation of all previous steps needed to transfer data into a relational spatial database facilitates almost real time visualization of collected data. For instance, Figure 5 illustrates a digital interactive web map showing the locations of Alternative Development projects. Data visualization and reporting of collected data is supported by modern open web technologies like PhP, JavaScript and the Leaflet mapping library. Applications like interactive maps or data collection dashboards have been designed at UNODC to retrieve information directly from the relational Database and publish to the web.

An important development for the project was the possibility to serve spatial data directly from the database stored in UNODC servers to dynamic web maps. The common approach of serving shapefiles or text files is popular, but when data is dynamically and continuously updated the automation of a synced database becomes relevant.

6. Users

A fundamental step in the implementation of a successful survey is user training. In UNODC enumerators have extensive experience in collecting data. However, not all of them were equally savvy in the use of Android systems, applications and digital content management. Several thematic trainings were held to explain the basics of the operation in the field and how to verify the correct configuration of the device as well as data transferring. Data analyst and survey coordinators were also trained on the use of the web visualization tools and dashboard of data collection campaign.

7. Discussion

This document presented the configuration of a tightly integrated system between data capture, data processing and visualization with an system suitable for the type of infrastructure present in developing countries. The system has been successfully deployed at UNODC for the monitoring and georreferencing of rural households participating in the Alternative Development program. The system relies on the ODK architecture for spatial data collection in disconnected mode and information technologies like scripting languages for data processing and web applications for visualization.

Data collection is, perhaps, the most crucial stage of the information cycle process. The opportunity to measure a field or interview a farmer is an irreplaceable event as the conditions enabling data collection are unique considering weather, financial resources, staff availability, field security and objects being measured. Teams long accustomed to conventional analogue data collection will initially be reluctant to the idea of a digital collection system, given the intangible nature of digital data as compared to conventional paper forms. To facilitate adoption of the system and reassure field enumerators at UNODC, we develop a dashboard on the website where they can login to trace the submitted data. The system provides them general indicators of their work such as the number and type of questionnaires effectively received, and a point map of the visited locations.

The deployed infrastructure is economic in terms of hardware and software resources. Most importantly, it is flexible and can be adapted to available budget. If for instance, a server is not locally available, data can be hosted permanently on third party servers for a low cost (e.g. Amazon, Google). Several GPS enabled Smarthphone devices can also selected depending on the local market availability. What is more crucial for the implementation of a full-fledged project as described here is to count with expert knowledge on information technologies. This is needed for database design and maintenance, server configuration for data collection, data scripting for data processing, and programming of web applications for visualization and data reports.

Fortunately local knowledge is growing rapidly in developing countries. Furthermore, there are generous developer communities in the Web around open technologies as the ones presented here. For instance, the ODK user community is moderated at the University of Washington and daily support is provided to ODK users, with a large volume of knowledge accumulated in the history of its forums. The same happens with the communities of R, PhP, Javascript and Leaflet. This shared knowledge adds a vital level of support that accelerates local knowledge and ensures sustainability.

8. Results

The impact of the implementation of the described system has enabled timely and accurate data collection and dissemination under a low cost implementation, using open source technologies. Firstly, data has been captured in an efficient way, with minimal data errors, minimal data loss and a more fresh transition between collected data to indicators and data visualization. Field enumerators have been relieved from carrying paper forms to the field and from memorizing questionnaires guidelines as the application guides the enumerators through the questionnaire restrictions and the logic flow.

Secondly, the information technology capacities of UNODC has grown rapidly with a group of information technology professionals that before spent most of their time scanning, editing and digitizing collected data, to a more rewarding and exciting environment where they develop skills in data manipulation, database management and data visualization tools. We consider this highly relevant, to close the digital divide between developed and developing nations, and ultimately, to promote a higher impact of information technologies in developing countries.

Figure 5. Digital interactive map showing field collected data directly stored in a database

REFERENCES

Dutta, S. (2009). *Global Information Technology Report 2008-2009*.

Hartung, C., Lerer, A., Anokwa, Y., Tseng, C., Brunette, W., & Borriello, G. (2010). *Open data kit: tools to build information services for developing regions.* Paper presented at the Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development.

International Telecommunication Union. (2015). Measuring the Information Society Report 2015

Nordling, J. (2015). GeoODK. from https://github.com/GeoODK