Reflection on data collection methods

Version 0.1

Traditional paper based methods has been the mainstay of data collection efforts by NGOs over many decades. Pre-printed blank formats are used by trained data collectors, be they NGO staff or volunteers, and the collected data is hand written into blank spaces in these forms. The forms are often photocopies of a printed out original form. The form itself is usually composed in a word processor (MS Word, Libreoffice Writer etc.) or spreadsheet (MS Excel, Libreoffice Calc) on a computer and printed out. This method is often used due to the flexibility it affords and the cost effectiveness. If the number of data collectors increases, then more copies of the data format need to be generated, which can be easily done either by photocopying or printing the forms.

While this method reduces dependence on computing devices in the field, there are a number of issues that arise in the subsequent stages. Paper forms that have been filled in need to be collated and managed so that the progress of data collection can be monitored. In places where the data collection area is spread out geographically, the paper formats need to be physically brought to a central location to be scrutinised. This may translate to an entire batch of data formats having similar errors before they could be corrected. It is not always feasible to go back to the respondents and correct the mistakes. In such cases, the errors would persist in the dataset reducing its overall usefullness.
Since it is difficult to restrict the data being entered in a form field on paper, there is scope for the data collector to fill data in an inappropriate form. It is also possible to miss out the units when entering quantitative data. It is possible to miss asking particular questions, or to skip pages entirely if they are stuck together. Descriptive questions may be answered in one or two words and quantitative questions may be answered in a descriptive manner. The nature of the medium allows an individual to override written instructions regarding particular questions, which would bring in differences in data collected by different individuals. Thus there are many possible sources of errors in the data collected using a paper based method.

The paper formats also need to be organised to ease entry of data into computers. Most often the data entry is done in spreadsheets, where good practices of data management may not necessarily be followed. Numbers and units may be entered together in cells, or columns merged. Colours may be used to indicate some information, or data divided into groups based on sheet or file names. This results in unanalysable data. Often a lot of manual effort would be needed to perform even simple analysis. Data entry itself is a laborious process and can introduce a number of errors if the data collector and data entry operator are different. If multiple persons are entering data into the computer, then in the absence of clear guidelines, they may each interpret the data differently and enter the data differently leading to inconsistencies within the dataset. In any case, the data entry process is labour intensive and time consuming.

Field conditions, particularly in rural areas, offer many challenges to data collection efforts. NGO staff and/or rural volunteers are often more comfortable with freehand writing than using computing

devices. Access to computing devices is also limited in most non-profits due to resource constraints. Many still rely on desktops for their work which do not offer any scope for mobility in the field. Where laptops are available, there may not be enough numbers to be sent out to the field for data collection. Moreover, battery backup, power availability, interface language, technical support etc. become constraints in the widespread use of these devices. For recording images and videos as part of these data collection efforts, digital cameras are needed. Where the location and/or elevation needs to be recorded, Global Positioning System (GPS) units are needed. Even in paper based data collection, other electronic devices are in use. The cost of acquisition and maintenance of these devices adds up over time and often it is difficult to replace them immediately, when operating in a project based funding mode. In the predominantly dusty, hot and wet field conditions in India, these devices may not last very long in the field.
There is thus scope for an alternate method of data collection where some, if not all of these shortcomings could be addressed. Until a few years ago, computing devices were expensive and were mostly in the form of laptops or desktops. Access to these were also mostly limited to a few people in most of the NGOs, due to resource, language and computer literacy constraints. The emergence of handheld devices such as smartphones and tablets has provided scope for their utilisation in data collection.

Recently, we have seen a surge in the ownership of android based smartphones and tabets, which are available from hundreds of brands with a broad range of hardware and software capabilities. These could be a basic smartphone from a no-name chinese manufacturer costing a few thousand rupees to flagship phones by market leaders costing more than a mid-range laptop. The key behind this range is the Android Operating System by Google, which is a mostly open source stack. Some of these devices lie in the so-called budget category (Price below Rs. 15,000) which offer a number of features that are of use to NGOs.

Given that these devices are cheap and available in most parts of the world, there are many different efforts to utilise these devices for various specific purposes by building apps (similar to software packages on Linux or Windows). One such app is Open Data Kit, a free and open source software that allows data collection through custom forms using any android device. It does not need internet in the field for data collection which is a rarity in the Indian countryside.

ODK can utilise the camera and GPS available in the device to record images and locations respectively. Forms once created and uploaded to a server can be downloaded to any android device which has the ODK Collect app installed in it. The ODK Collect app works with almost all the versions of Android in the market and does not require anything beyond a basic hardware configuration. This provides flexibility in terms of investing in low cost hardware or utilising android devices owned by staf or volunteers for the data collection. The devices do not need internet connectivity for data collection and therefore there is no need to have decidated SIMs in each device. If a Wifi network is available, then that is sufficient for data upload to the server. It is also possible to configure the systems so as to avoid connecting to the Internet at all and locally sync the data to a server on a Local Area Network.

There is a need for the design of data entry forms using xlsform or other tools which requires some knowledge of working with database concepts. There is the cost of acquiring the devices which can be justified in case of repeated data collection over time or across various surveys. The advantage offered by this method is that form fields can be tightly defined.E.g. if a field takes numerical data then text cannot be entered there. If, then, else logic can be enforced to ensure that questions are properly administered. Images and GPS locations can be automatically recorded as part of the appropriate records reducing the effort of manual linking of these which may lead to errors and are certainly time consuming.

Perhaps the most important advantage of this method is the lack of manual intervention in importing the data in a form that is amenable to analysis, which reduces the time between data collection and data analysis. ODK Aggregate, which is a tool to import the data from the devices, has built in data visualisation tools that can map the locations of individual records and also create pie/bar charts for each variable. This can be used to quickly check the data quality and isolate outliers.

Since most of the Android devices are not meant for rough handling in the field, it would be advisable to invest in screen guards, carry cases and power banks, which are all now commonly available. Instead of buying the latest model, going for a model that is a year or so old would be more economical without any major compromise in terms of performance.

There is plenty of documentation online including tutorials on how to go about deploying an ODK based data collection system. The Android interface is quite intuitive and the learning curve is minimal when it comes to using an ODK form. If the devices are pre-loaded with the form, then entering data is quite straight forward. As with all data collection efforts, it is advisable to train data collectors and to check the quality of the data collected by them before deploying them in the field.

Investing in Android and ODK based system ensures that our data remains accessible even in future as these are open software. ODK allows data to be exported as comma separated value (CSV) files which can be imported by any spreadsheet software such as the FOSS libreoffice or proprietary MS Excel. Especially for non profits that use public funds for their work, it is important to invest in open standards and formats so as not to get locked in with particular vendors.