**UNGIWG Task Group 5**
***Global Navigation Satellite Systems***

November 2007 Progress Report

# Embedding data quality in "GPS-tagged" field data surveys.

Patrick McKay – WFP Johannesburg Regional Bureau
George Mu'ammar – WFP HQ - Rome

## *1. The issues of field Data collection*

The Vulnerability Analysis and Mapping branch (VAM) in WFP is systematically engaged in various primary data collection activities using a standardized approach. However primary data collection mechanisms can be time consuming. The traditional processes of data collection, geo-referencing, data entry, processing/cleaning, analysis and report preparations require heavy investment of time and training and yet data quality issues are continually present. Speedy and geo-referenced data capture is a growing and critical bottleneck in the assembly of VAM datasets. This is impacting on the timeliness of undertaking the relevant analysis. A recent review of VAM activities shows a big time lag between data collection and reporting the results.

The problems of data quality and consistency in traditional surveys are mainly due to the separation of the data collection activity from the data entry activity. Enumerators write the results of their interviews on paper questionnaires, and at the end of the data collection process, these will be handed to a data entry team who will enter the data into a database. They will discover that the responses taken have not necessarily respected the nature of the questions and the logic of the questionnaire. Some questions should be skipped based on the response of previous questions, and this may not have been correctly interpreted by the enumerator. The answer to some questions consists of a code to be read from a look-up table, which was memorised by the enumerator, and in some cases the wrong code was written. Validation that is taking place during the data entry fails on some of the responses, whose value should fit within a domain range, or in the case of proportional piling may not sum correctly to the total. These are data collection errors introduced by the enumerator while they are concentrating on the interview with the household, which requires some attention.

Not always do the paper questionnaires reach the destination dry and in proper order. The data entry team in reading the results and performing the data entry into the database may introduce further errors due to misinterpretation of the handwritten responses, and may not know how to cope with mistakes that make validation fail during the data entry process, since the mistake has already been made and it is impossible to return to the data source.

Other issues related to field data collection follow:

- Preparation of the questionnaires for the survey. Tens of thousands of printed pages of paper are required for the data collection in any medium to large survey, and an ad-hoc database and front-end data-entry system needs to be

developed for the data entry phase.

- The location aspect is probably the most disturbing item to manage amongst the questionnaire's fields. All questionnaires request village name, code and administrative unit names in a hierarchical manner. However a certain number of responses will not be properly located due to ambiguities and inconsistencies in the location names. Garmin handheld gps units have been used but results were not fully satisfactory due to complications in coordinate formats and cable connections for download of waypoints.

- A survey is based on a sampling frame. This is usually defined on a village basis, but can also be defined on a geographic grid created in a GIS. Once in the field each enumerator must know in which sampling unit they are in or how to locate the village they are assigned.

# *2. "GPS-tagged" PDA data collection*

## 2.1 Principle

The envisaged solution to the problems listed above is based on merging data collection and data entry into a single process. The tool identified to best suit this requirement is the PDA (personal digital assistant) i.e. a handheld computer with a touch-screen. Today PDAs are used commercially for location-based services and for data capture in the field, often in combination with a GPS and a digital camera. All models have inbuilt GPS or wireless connection to an external receiver. Although its computing power is very limited, it is sufficient for data entry requirements.

WFP VAM has developed a PDA based application for household surveys. It allows rapid data entry into a questionnaire using drop-down select boxes and other typical components of a data entry form. It also implements *rosters*, i.e. a table of questions with a variable number of rows. The software connects to a GPS through a wireless Bluetooth connection to acquire location coordinates and date and time of usage.

The software requires a questionnaire definition to run. The questionnaire is converted into an XML file during the preparation phase, which requires one working day to create and test. The questionnaire definition contains all the field definitions and the logic of the questionnaire (validation, field skips, question dependencies and rosters).

The GPS data collected is read directly from the NMEA 0183 string and is saved with full decimal accuracy. The coordinates are latitude and longitude decimal degrees and refer to the WGS84 datum. Together with the coordinates also the date and time (as received by the GPS) and the HDOP (*horizontal dilution of precision* – an indicator of accuracy) is saved.

In terms of hardware, the external Bluetooth GPS receiver was identified as an ideal solution mainly because of its power independence. A GPS receiver's power consumption is of comparable order to the power consumption of a PDA, even when properly configured for minimal power consumption, and therefore a PDA with internal GPS unit will give 4-5.5 hours of power autonomy, whereas a PDA with an

active Bluetooth connection can run for more than 8 hours with a single battery charge. The external GPS unit has its own inbuilt battery and power switch, allowing it to be switched on only when needed and therefore prolonging the duration of a single battery charge to a few days' usage.

In terms of data security, the survey software saves the data on the PDA's internal memory and also on a removable memory card too (generally an SD – Secure Digital card) which has become the standard removable media for all PDAs. The SD card is a tiny non-volatile memory chip which can be easily removed from the PDA. The smallest SD card available today of 512 MB has a retail price of 10 USD, and provides far more capacity than the requirements of the survey. The size and ease of use of the SD cards allow them to be distributed at the start of the working day and collected at the end, thus enabling the data to be stored and transported separately from the electronic devices.

The use of SD cards has also facilitated transfer of data and software between PC and PDA. Through the use of an SD card reader, the SD card becomes equivalent to a USB Pen Disk, and therefore there is no need to use the synchronization cradle and software that comes with the PDA. The survey software developed by WFP includes an autorun procedure for the SD card, which automatically installs the software and questionnaire definitions upon insertion of the card and uninstalls them upon removal of the card.

## 2.2 Benefits

There are two main fundamental benefits to PDA based surveys. These are data quality and reduction of costs of the survey.

Data quality is improved thanks to the following features:

- Fields flagged as mandatory must be completed before the interview can be finalised.
- Field skips are automatically handled.
- Look-up tables are implemented as drop-down boxes containing both code and description.
- Responses are validated through the use of domain range of values (to exclude accidental ridiculous responses) and proportional piling (numbers must add up to 100% for example). Text fields are limited by length.
- Exclusion of items from drop-down boxes, to ensure that the same response cannot be selected twice.
- The type of data requested for each field is enforced (integer, text, date).

Costs of the survey are reduced. Clearly there are overhead costs of purchase of the PDA, GPS units, SD cards and car power charger units (for countries where electricity distribution is limited).

A practical example of the costs involved in preparations of a survey with 40 enumerators completing 2000 interviews:

| **PDA Solution** | **Paper Solution** |
|---|---|
| PDAs $16 000 ($400 each, re-usable for many surveys) | Preparation paper questionnaires $500 |
| Database $0 | Database $2 500 |
| Data Entry: N/A | Data Entry $3 500 ($35/day for 20 qs) |
| Data Cleaning: N/A | Data Cleaning $1 500 ($150/day for 10 days) |
| **Total = $16 000** | **Total = $8 000** |
| Cost per questionnaire = $0.80 (assuming PDAs will be re-used for 10 surveys) | Cost per questionnaire = $4.00 |
| Data preparation Time: 1 day | Data preparation Time: 30+ days |

As well as reducing costs, also the time to availability of the data is reduced. Although the enumerator training requires 2 or 3 additional days for training in the use of PDA and GPS (depending on the team's computer literacy and familiarity with smartphones), all the time for data entry after the data collection is eliminated (usually in the order of 100 workdays) and data cleaning is no longer required, or at most is minimal.

The performance of waypoint collection is greatly enhanced through the use of PDAs, mainly due to the fact that the user does not have to deal with a non-standard interface of a handheld unit such as a Garmin. The PDA can inform the user through natural language messages as to the errors or lack of precision of the GPS at that time. The association of GPS coordinates to other data is seamless, since the GPS coordinates can now be embedded into a larger data collection frame, as opposed to having to deal with capturing coordinates separately from the other data.

Contrary to common belief, data security can be improved with the use of PDA compared to using paper. The fact that the data is saved on the SD card as well as on the PDA means that there is a non-volatile backup of the data which can immediately be copied. During a survey, when the enumerators regroup at the end of the working day, the team leader may collect the SD cards of the group and copy all files to their PDA and then onto a separate SD card in a matter of minutes. This form of immediate copying and backing up of data is impossible in the field when dealing with paper. The SD card is non-volatile and will resist exposure to percussion, to water, sand and other hazards.

Another issue, which may depend on the country and the agencies involved in the survey, is that of trust. The work of an enumerator is quite tasking and some may be tempted to not actually go around to each sampling unit designated and locate households willing to reply to the interview, and would rather use their own imagination as a data source. The seamless use of a GPS enables the time and location

to be monitored and stored within the dataset when the data is stored and this stands as a form of proof of the time and location of each interview.

It is needless to be said that use of digital media reduces the waste of paper, not only in terms of costs but also in terms of environmental impact. A survey of 2000 households with an average questionnaire of 10 pages, even if printed double-side, still means that at least 10,000 sheets will be used.

Last, but certainly not least, is the building or capacity and provision of modern technologies. The next section will look at this issue.

### 2.3 The role of capacity building

Capacity building, or capacity development, is a fundamental aspect of WFP and of all UN agencies that operate in developing countries and in fact many organisations are specifically mandated to develop technical and technological capacities. In WFP household surveys are nearly always performed in collaboration with the government and other UN agencies and NGO partners. The participation of local enumerators is fundamental because of their knowledge regarding the geography and the demography, the customs and the issues regarding the ethnic diversities.
During the past years WFP has trained many hundreds of enumerators, mainly in African countries, in the use of PDAs for data entry. This training was embedded in the enumerator training which also analysed the questionnaire's questions and logic, and was then followed by several weeks of data collection to put into practice what has been learnt.
In most cases the training was well received and provided a boost in motivation for the enumerators. In some cases the survey took up to 30% less time than expected, due to facilitated data entry and also to higher motivation levels due to the technology being used.
The use of Bluetooth GPS units provided a further level of interest, since these devices can also be linked to common smartphones and put to use with easily available navigation software.

## *3. Use of PDAs within WFP surveys*

For the period of 2004-2006, the use of PDAs was limited to the southern Africa region (ODJ), and was managed by one consultant capable of programming the PDAs. In this time approximately 30 surveys were conducted including a number of Joint assessments together with UNHCR, a number of baseline studies and many other surveys. These were conducted in 9 different countries using English, French and Portuguese.

In 2007, the joint ODJ/HQ initiative began, with the aim of automating the process and replicating the use of PDAs to other regions. In order to meet the increase in demand for PDA based surveys, the questionnaires would not made into a programme, but would be converted into a set of instructions onto the PDA, which would take the form of an XML file which the PDA would interpret and then ask the questions specified. Also the GPS component has been added at this stage.
During this year surveys have been carried out in numerous African countries and in the Caribbean using the client version of the PDA software. The questionnaires have

been prepared in a few hours of work and quickly tested. Response has been positive in nearly all cases, and all offices request to retain a number of PDAs for continual updating of survey results.

The final step of the process is to distribute the tools necessary to enable non-technical staff to handle PDA questionnaire creation. A questionnaire creator based on MS Access is already in use, and at a recent training, all 18 participants were able to use it to generate a simple questionnaire. An online system is being designed which would simplify further, and guide the user in the process of creating the PDA instruction set.

## *4. Data collection in the context of a UNSDI*

Since version 1, the GeoNetwork software was originally developed by FAO and WFP VAM, and later with participation of UNEP and today also by the open-source community. WFP VAM today has a network of 10 distributed nodes of GeoNetwork within its Spatial Infrastructure Architecture (VAM SIE) for cataloguing, sharing and publication of geographic data.

Within this framework WFP VAM is currently enhancing GeoNetwork functionalities to include online uploading of data collected in PDA based surveys. The database and data upload function (in beta phase) is currently being used online at the Johannesburg regional office of WFP.

The database being developed is designed with a flexible schema to be compatible with any data structure and therefore will be independent of the questionnaire structure.

The current version of the PDA survey application captures point locations. Even though there is no limit to the number of points that can be captured per single questionnaire, in the case of household surveys only one point is needed. In these cases the data is automatically georeferenced, since all alphanumerical data can be attributed to a single point in space and in time.

Very recently collaboration with UNJLC is taking place at the time of this writing to develop PDA based forms for road assessment and obstacle reporting used in defining routes for logistics operations. Road assessment is being captured by reports on the status of road segments defined by a start and end waypoint. These waypoints will be overlaid to the GPS track of the road in order to define the segments to which the reported status will be attributed.

## *5. Future developments*

Future developments are envisaged mainly in 2 areas namely enhancing the GPS related functionalities and the development of online facilities.

Several applications have been proposed and are currently under development based on the requirements to capture line and area features from GPS reception. Line features will be captured as a sequence of points either taken individually or based on

a GPS track i.e. a continuous stream of points that will be filtered using a time interval or a minimum distance interval between each point.

Area features will be captured using the line feature principle to define a closed polygon, which, once tested for topological correctness (in case the polygon is self-intersecting) will have area and perimeter calculations performed directly on the PDA.

Although accessibility to online services remains limited in countries where household surveys for food security assessments are taking place, several aspects of the project are being designed for future development. These consist in the creation of an online repository for questionnaires, and an online wizard for the rapid creation of questionnaires. These two features will be integrated in order to provide a minimal level of standardisation and reusability of questionnaire sections and questions, and also in the access and analysis of data. Online collaboration tools will be included to facilitate collaborative authoring of questionnaires, in order to provide strong versioning and history of the questionnaire editing.