ELECTRONIC DATA COLLECTION: MORE QUALITY, LESS CLEANING!

This publication has been produced with the assistance of the Office of the United Nations High Commissioner for Refugees (UNHCR). The contents of this publication are the sole responsibility of CartONG and can in no way be taken to reflect the views of UNHCR.

1. Introduction

It is highly recommended to spend the necessary time before a mobile data collection deployment to ensure that all the possible data quality checks have been implemented, to save time, energy and gain in the quality of the data collected. The purpose of this document is to serve as checklist providing you with the necessary reminders, on the technical and methodological aspects, to keep data cleaning to a minimum! Some of the elements below may seem obvious, but hopefully having them all available in one place can help ensure that none falls through the cracks!

Many technical references are made to XLSForm that is the most commonly used form conception format in the humanitarian and development sector.

2. Methodological design of a MDC form Checklist

Note that only aspects directly linked to mobile data collection are mentioned here- Don't hesitate to also read the «Questionnaire Design for Needs Assessments in Humanitarian Emergencies, Summary » accessible here : https://www.acaps.org/sites/acaps/files/resources/files/acaps_questionnaire_design_summary_july_2016.pdf, for more general data collection conception principles.

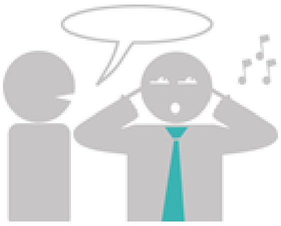
Ensure you do not lose anything in terms of "Do no harm" principles by using electronic data collection

- Examine whether **using mobile devices will not create data protection issues** in your context – i.e. using mobile devices during interviews with children or vulnerable populations can be questionable ethically, but there are many places where they can also be seen as tracking tools (i.e. Syrian context, Khost in Afghanistan…),
- Ensure the **consent** you ask of respondents (for a survey) **includes the fact that a mobile device will be used**, giving the possibility for an enumerator to also use paper rather than mobile in case the interviewee refuses,
- Check that the different tools that you use include the security features (such as encryption, relevant access rights etc.) that are deemed necessary

Interesting links to check out:

- *Responsible data hackpad with many useful links:* https://paper.dropbox.com/doc/Responsible-Data-Hackpad--AP9Z9oVaTB9NnAHt6eTbZt5CAg-SA6kouQ4PL3SOVa8GnMEY
- *Security in a box website :* https://securityinabox.org/en/
- *Secure messaging apps comparison:* https://www.securemessagingapps.com/
- *Responsible Data handbook:* https://responsibledata.io/resources/handbook/
- *WFP manual on Conducting Mobile Surveys Responsibly:* https://documents.wfp.org/stellent/groups/public/documents/manual_guide_proced/wfp292067.pdf
- *Cash related data starter kit for humanitarian field staff:* http://elan.cashlearning.org/
- *Data protection starter kit developed by CartONG & Terre des hommes:* https://www.mdc-toolkit.org/data-protection-starter-kit/
- *Security settings suggested by the ODK community:* https://docs.opendatakit.org/collect-security/

Choose a relevant app, device, interface to do the job!



- **Ensure using electronic data collection is acceptable** in the context of intervention (for cultural or security reasons) and also that it is the right method to implement (for unstructured qualitative discussions it would not be)
- Check the **mobile device has the features** you might need, such as functional GPS working offline, decent camera, good screen space, ruggedised…
- Check the **app has the necessary features** (follow up of an entity, good encryption, being able to scan a barcode, working fully offline or just for the data collection…)
- Keep in mind that somebody out there has probably already **devised a similar form** that will avoid you completely reinventing the wheel
- **Be wary of the magical harmonised or imposed app** suggested that might not be adapted to your specific context. Check the **interface** you will use for collection (mobile and/or webform) is **well thought out**. Make sure you adapt the design to the interface in question to make it as user friendly as possible and also because there may be features with different functioning modes between these different interfaces.

Interesting links to check out:

- *Comparison of features available in ODK and Enketo webforms:* https://xlsform.org/en/ref-table/
- *Comparison of a certain number of tablets and smartphones:* http://blog.cartong.org/2018/01/18/mobile-data-collection-in-emergencies-which-phones-or-tablets-can-do-the-job/
- *Blog post on helping figure out the best mobile device:* http://blog.cartong.org/2017/04/27/choosing-mobile-device/
- *Benchmarking of 17 different MDC solutions:* http://blog.cartong.org/2017/08/14/mdc-benchmarking-2017/
- *Benchmarking of different solutions for MDC with a strong geographical component:* http://blog.cartong.org/2017/05/05/benchmarking-mdc-tools-with-strong-gis-component/

Adapt your electronic survey conception to your data analysis plan

- Make sure that your **exported data is as "ready-to-use"** in your given data analysis context as possible.
- Consider any **multimedia** you might want to include that might be relevant **for your operational understanding, quality control or accountability/reporting** (adding a photo of the infrastructure to double check it's categorisation, a GPS point, an audio recording of the interviewee, a drawing…)
- Apply exhaustive **data validation logic**: there should never be open numeric fields (maximum and minimum constraints can nearly always be derived from secondary data and/or piloting/experience), you can also use Regex constraints, limit the number of entered characters for text fields if it makes sense, ensure that multiple option answers selected are adapted (ie that you have selected a minimum or maximum number of options, that you have not selected "none" at the same time as another option etc), that a date selected is in the right range (past or future) etc.
- Make **mandatory** as many questions as you can- consider however only questions that can be filled in in 100% of cases (e. g. do not make a GPS point mandatory as it depends on the stability of your mobile device for this feature)
- Verify that the mandatory **single-choice/multiple option questions all have the necessary safeguards** ("Don't know", "Unknown", "Other", "N/A", "None of the above", "Refuse to answer", etc.) to avoid forcing your enumerator to enter incorrect information to be able to submit his data.
- Make sure you use **multiple answer questions only for questions where it is really makes sense** and that you are sure you are in capacity to analyse (as they can be more complex to analyse than single option questions). Consider using "ranking" questions that will give you a "pondered" understanding of option responses such as "What is your primary source of water"/"What is your secondary source of water"...)
- Set up the necessary **calculations for your indicators in the form directly** to avoid having to set them up later in the analysis tools
- Reflect on ways to adequately **triangulate data on some key indicators** to double check its relevance (i.e. showing the result of calculations to the enumerator directly in the field and asking him for confirmation), adding alert messages in case of possibly inconsistent data

Interesting links to check out:

- *Regex expressions validator: https://regexr.com/*
- *Blog post on using Regex contraints: http://blog.cartong.org/2018/03/15/all-about-regex-in-xls-form-when-how-and-examples-in-the-humanitarian-and-development-fields/*
- *Blog post on advanced XLSForm coding part 1 : http://blog.cartong.org/2015/08/11/advanced-xls-forms-coding-1/*
- *Blog post on advanced XLSForm coding part 2 :*
- *http://blog.cartong.org/2015/08/21/advanced-xls-forms-coding-2/*

Closing the loop: Ensuring data capture is short and sweet and keeping in mind that "there is no such thing as common sense"

- **Identify all your free text questions and check whether it might be relevant to have instead a possible list of options** with the "other" safeguard (administrative information, name of enumerators…) - Plan preparatory work such as Focus Groups

Discussions if in doubt concerning the possible list of answers in that given context - Avoid open-ended questions in particular for data that you want to analyse quantitatively or for disaggregation purposes (and also so that your enumerator will gain time filling in his submissions and it will reduce interviewee fatigue) - Keep in mind that sometimes using combined qualitative and quantitative methods can be very useful, but it might not be relevant to combine the methods in electronic data collection

- **Group questions and put section titles as well as use colour in your labels** to facilitate visual understanding by the enumerator
- **Add any necessary tips** (hints, customised constraint messages, audio explanations) concerning all the difficult definitions/ notions/ jargon/ acronyms/ measurement units, or provide an enumerator's guide if necessary
- Check that your **appearance settings are compatible** with your device size (grouping questions on the same screen or using big matrix-type questions when you have a small screen is not recommended) and your question meaning (using likert scales for questions where it does not make sense)
- **Translate the form** into all necessary languages to avoid mistranslations by enumerators.

  See if some **variables could not be set up as cascading lists** to avoid any error entries and ensure than any list of options is as short and relevant as it can be- any extra click is a possible error!
- **Consider whether there might not be data that you have available in another database** that you could either avoid collecting in this survey or reinject directly in your submissions through calculated values using "external CSV" features to reduce errors and limit data entry

Interesting links to check out:

- *Check out the following document on the https://www.mdc-toolkit.org/design-your-forms/ page: "Managing external data", "Adding media as responses or helpful notes", "Working with different languages", etc.*

Conform to basic data management principles- you won't regret it in the long run!

- Give your survey a **title** that speaks for itself and that you are sure will not create misunderstandings for enumerators (with the relevant information such as thematic component, location, year, etc.)
- Include the necessary project **metadata** (recording the start date and time for the submission, including intermediary time stamps, IMEI number of the phone used, etc.)
- Look into constituting a **unique identifier** for your dataset if it is relevant to your project and see how best to have it captured (calculated automatically based on other information, selected in a list, filled in manually with advanced constraints, scanned from a barcode or QR code, etc.)
- If you are using administrative data, integrate standard and interoperable **P-codes** (place codes) that will be understandable if you need to link your dataset with other past or future data collections (link to HR post on the question)
- Set up relevant Standard Operational Procedures concerning the **versioning of your form** to follow its life cycle and ensure data consistency across different version (and don't hesitate to archive form versions and data accordingly),
- Prepare a Standard Operational Procedures concerning your **"data quality" schedule/plan** during the data collection (if you wait until the end of your data collection, it may just be too late).

Interesting links to check out*:*

- *Integrating P-Codes in XLSForm:* *https://www.humanitarianresponse.info/fr/applications/kobotoolbox/document/how-use-p-codes-kobo*

Testing testing testing and more testing!

- Make sure you have the tools **tested by your thematic project managers, technical colleagues familiar with mobile data collection, enumerators,** and any other relevant stakeholder - they may have some good ideas for improvement that you may not have thought of.
- The most important aspect to test will actually be that **test/pilot data** actually corresponds to what you expect it to correspond to and **is usable in your analysis tools / compatible with your analysis plan**. The pilot should be a close to real conditions as possible: same mobile devices, same context, same type of interviewees, etc.

3. XLSform design technical checklist

- **Respect good practices for the naming of your variables and options** ("**name**" columns):
    - Be consistent in naming to facilitate your analysis and the use of your form in other contexts,
    - specify without any space/special characters, ensuring also that they do not start with numbers
    - avoid separators that might cause issues with your chosen analysis tool inside your variable and option name
    - Avoid any duplicate names of variables
- If you use **groups or repeats** in your form, **ensure you have a pair** for each one.
- **Avoid blank lines** within your form, can cause problems on some analysis tools, **and unnecessary spaces in the XLSForm syntax** in columns other than question and answer labels
- **Filter your "Survey" tab column by type of question to ensure all relevant settings are present** -and test each parameter you have set up on the different interfaces you plan to use (webform, mobile application, etc.):
    - constraints (i.e. all integer type questions having a minimum and optionally a maximum, having Regex constraints for questions for which it makes sense, etc. - see recommendations in section 1)
    - skip patterns,
    - calculations,
    - types of appearances,
    - mandatory settings for integer, decimal, select one, select multiple, date, text types of variables.
    - common hint settings (i.e. all multiple option questions having a "please tick all that apply" hint to avoid comprehension errors)
    - languages are filled for each question/ list of choices (keep in mind that you can also ensure hints, constraint messages, mandatory messages, etc. are available in the different languages of your survey if you want to go to those lengths)
- Create an **automatic naming** of your submissions to facilitate the follow-up of submissions by your enumerators ("instance_name" column in your Settings tab)

- Ensure you have a **"form_title", "form_id", "default_language"** in your setting tab to facilitate understanding of your form and usage in the field

Interesting links to check out:

- *Check out the following document on the https://www.mdc-toolkit.org/design-your-forms/ page: "Recommendations for the naming of questions variables and answer choices"*

4. Logistical recommendations

A few logistical reminders can also be made to avoid bad surprises in the data quality or to avoid field issues:

- **Deploy from a clean slate**: Empty all previous test or real data / forms from the phone to avoid "noise" or an enumerator selecting the wrong form or wrong form version, delete any unnecessary apps and multimedia to gain space.
- **Update your phones' general settings** (date and time, language, time for screen saver mode, etc.) **and the apps you mean to use** to avoid manipulation or technical issues
- Check the **GPS precision** of your devices if required, as well as **camera quality** (if you are scanning barcodes, QR codes or need high quality pictures…)
- Print **paper back-ups,** user guides or paper help sheets when necessary (HH composition, etc.)
- Clarify the **roles and responsibilities** related to the electronic data collection and general life cycle of the submissions (who validates, who sends, etc.) so that the procedures are well determined
- **Print relevant** paper version of the electronic forms and all other **supporting material needed for field enumerators to get the electronic data collection job done** in the best conditions possible (HH composition forms, etc.).

Interesting links to check out:

- *to improve location: https://docs.opendatakit.org/collect-location/*

For further reading:

- CartONG blog : http://blog.cartong.org/
- MDC Toolkit: https://www.mdc-toolkit.org/mdc-guidance-and-documentation/
- ODK forum and guidelines : https://docs.opendatakit.org/collect-best-practices/ as well as: https://forum.opendatakit.org/
- XLSForms website: http://xlsform.org/ and
- XForms specifications: https://opendatakit.github.io/xforms-spec