# THE NEW NORMAL: COLLECTING DATA AMIDST A GLOBAL PANDEMIC

Author: Wilhelmina Welsch, JIPS
May 2020

With the onset of the global COVID-19 pandemic and the efforts to contain it through public health measures, social distancing, lockdown and quarantine measures, the data and assessment landscape is facing long-term impact which could deliver the overdue push for all stakeholders providing international assistance to re-invent not only how to collect data in the coming months but also how to make use of other types of data that is readily available beyond the classic needs assessment/survey data. The crisis provides a clear window of opportunity to not only strive for creative and novel approaches to respond to its challenges but to finally translate the accumulated rich experience into current practices.
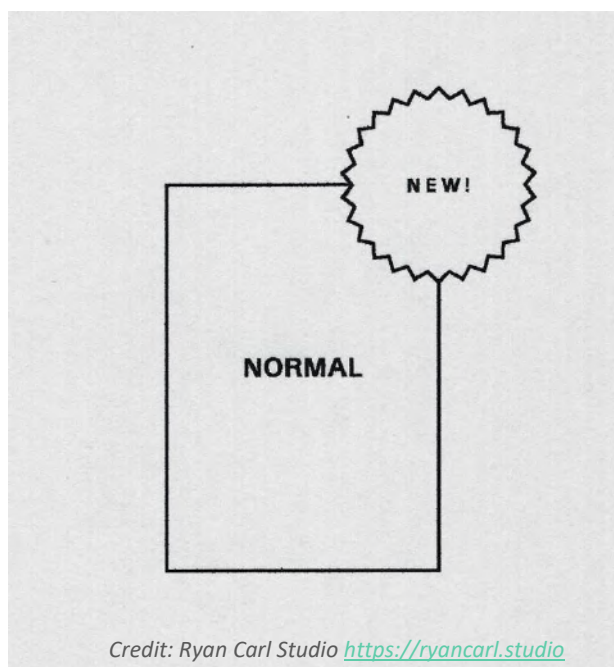
Credit: Ryan Carl Studio https://ryancarl.studio

The ultimate soul-searching as a collective will force us to get to the bottom of questions we have been discussing on both technical and policy level profusely over the last five years: how to be more effective *and* ethical with regards to data collection and analysis.

The challenges that the pandemic brings to accessing communities and collecting data will not only impact the humanitarian community but all stakeholders relying on up-to-date data for response and planning, including national governments, international and national organisations, as well as other service providers such as actors of the health and welfare systems. The humanitarian community is uniquely placed, however, to bring in their experience in operating in situations with limited or no access at all.

This article is intended as the beginning of a conversation between technical experts, and not as a guidance document. It thus does not aim to provide a comprehensive discussion of all possible methods, approaches and best practices. It rather presents a set of core assumptions that will have to

be tested at a later stage for the establishment of related policies, and ultimately response, to the COVID-19 pandemic.

1. Adapting data collection processes

There is an opportunity to adapt and look to those who are relying on vast amounts of data collected remotely: market researchers, labour force statisticians among others, as well as Mobile Network Operators (MNOs). There will be an even more urgent need to work much closer with affected communities and reflect their realities with regards to access and use of information. Similarly, there will be a need to link up with the scientific community, especially from social sciences and econometrics, as well as with modelers and demographers. There is a body of research that remains below the radar, which showcases significant progress in better understanding the explanatory power and characteristics of survey methods such as phone surveys or self-administered online surveys – beyond the every-day tools and methods used by most of us for sample-based household surveys, which we rarely fundamentally update or question.

Producing representative, disaggregated, agreed-upon, and up-to-date data is challenging. Producing representative, disaggregated and agreed-upon data on internal displacement is complex in protracted crises and even more so in emergencies. Our review of sampling methods in IDP contexts (to be launched end of May) highlights the difficulty of conducting sample-based surveys in contexts where dynamic factors influence the operational reality and where target groups are difficult to find and sample for. These issues will become even more pertinent in a remotely-management environment with little direct access to those populations.

All of the above highlights why thinking about adapting our processes requires both rigor and the adoption of new practices.

2. Short- and long-term shifts for humanitarian data collection

Given the above, there are a number of immediate as well as long-term shifts that will determine how we will work:

a. **Lean Design:** We will have to think *a lot* more strategically about **design** and **analysis** of the questions we need answered.
b. **Definition of Secondary Data:** We will have to re-think the definition of and tools for analysis of secondary data **– GitHub not Reuters.**
c. **Act as a Documentarian:** We will have to act as documentarians with regards to our approaches, and even more importantly with regards to the challenges and gaps we face**.** Reducing bias, documenting bias, and being cognizant of who is being left out being a key element.
d. **Shifting from Face-to-Face to Remote Technologies:** We will have to shift from face-to-face to remote technologies – where mixed methods will remain key. And we cannot rush the shift.
e. **Focus on Community Engagement:** We will have to emphasize community engagement and capitalise on thick data.

Let me unpack those in the following sections.

a. Lean Design

We will have to think a lot more strategically about design and analysis of the questions we need answered.

Two major determinants of this crisis will require us to shift our thinking with regards to data needed and analysis produced:

Our **inability to conduct research face-to-face**[1] during the pandemic, which allows for the collection of more complex information, will require us to condense and simplify what questions to ask of respondents focusing on surveys where they add most value:

"Surveys are particularly useful for collecting information about what people are thinking or doing now. This type of information is generally not available in most databases"
(Dillman, Smyth and Christian, The Tailored Design Method, 2014).

While it can be argued that with the use of finely granular data on mobility, financial transactions, etc. we could collect information about what people are doing – Sandy Petland from the MIT Media Lab's Human Mobility Department said in a workshop on Big Data I attended in Boston in 2018 that never before have social sciences been able to observe social interactions in real time[2] – the expression of priority needs, intentions, and plans as well as coping are best understood when asked directly from those affected.

What this current crisis is further showcasing, is the need to **think beyond the limited scope of humanitarian needs assessments** for analysis and take into account **non-traditional data sources** able to provide the bigger picture in terms of long-term effect of the pandemic as well as its anticipated needs.

Data on global supply chains will be one key element – sensitive enough to already provide pointers on anticipated shortages as well as price increases and groups who will be excluded from access to goods and services.

While mobility data will remain an important element, it will only be one element of what we need to know. What we will need to know, is much bigger in scope due to its domino effects and will need to be understood in a truly inter-sectoral manner: we need to understand context rather than just the situation of people in isolation. This means moving way beyond the limitations of siloed sectoral thinking and truly implementing the triple nexus, with a focus on functioning systemically and inter-

---

[1] Advice from health experts are that face-to-face be avoided at all costs – meaning, only extremely essential information should be collected, otherwise it should wait until after the pandemic. If we advocate face-to-face with social distancing, we are going against the generic 'stay at home' guidance in place in much of the world where outbreaks are ongoing. **Bound to "Do No Harm" principles, we have an obligation to adhere to the best practice model in this regard rather than pushing for an 'anyway we can find to fit within the parameters' approach.**

[2] See also Barratt and Maddox, (2016), "Active Engagement with Stigmatised Communities through Digital Ethnography: We anticipate that further unobtrusive analysis of digital trace data in this environment will be able to draw on the findings of this study in order to contextualize and interpret the social and behavioral patterns observed".

connectedly as well as developing and testing indicators that can be used to showcase living standards, deprivation and adaptation from a much wider angle.[3]

While some actors have begun shifting towards preparing phone surveys[4], complex structured questionnaires currently used in the humanitarian and development domain will need to be adapted significantly to fit within a 15-20-minute acceptable timeframe for the interview. And this is not to say that organisations have not been working towards adapting even complex and very structured surveys to new modes in the sector already. Many good examples outlined below (see example box in From Face-to-Face to Remote Technologies section) have been tackling this transition.

This is not only a challenge but also a huge opportunity to review our practice, which has been one of over-assessing populations for quite some time, ultimately reducing the need to collect data as a whole system.

RULE OF THUMB

The **rule of thumb** in this pandemic should be: collect remotely; if you can't, think of why this information is essential and provide arguments for why you can't get it any other way.

Choosing the right unit of analysis is a key part of those considerations of purposefulness and effectiveness of slim analysis. This includes questions like: when does household-level information not work? When do self-administered studies?

There are a series of thematic areas that will need to be covered on individual level, particularly with regards to health and protection as well as intentions - this is something we already know from household-level data collection that will become even more pertinent in the remote data collection context.

This needs careful consideration in the choice of unit of analysis as well as the choice of respondents and composition of any panel.

Using tailored designs that are customised to the survey population, topic, burden, and contact possibilities as well as the budget is seen to be more effective than attempting to use the same data collection procedures for all situations (and with it also the same panel for all types of questions).

---

[3] See also ILO: This requires us to widen our focus beyond the very limited number of key indicators, which are often a primary focus of labor market analysis. These indicators, such as employment and unemployment, will remain critical, but will be insufficient to fully describe the impact of COVID-19 on the labor market, workers and their households. This is not a failing of the concepts of employment and unemployment, rather a re-emphasis of something evident, the labor market is too complicated and diverse, as well as changing too rapidly, to be summarised into two statistics. A new, wider focus can be built into existing labor market statistics publications or additional publications to ensure wide dissemination and as comprehensive an understanding as possible of the impacts of COVID-19 on this aspect of life and the economy. https://ilostat.ilo.org/topics/covid-19/covid-19-impact-on-labor-market-statistics/

[4] See World Bank High Frequency Mobile Phone Surveys of Households to Assess the Impacts of COVID-19: http://documents.worldbank.org/curated/en/docsearch/report/148213

b. Definition of Secondary Data

*We will have to re-think the definition of and tools for analysis of secondary data:
GitHub[5] not Reuters*

As historically in any other crisis, particularly at its onset, actors are forced to rely on available secondary data to understand the bigger picture and plan for primary data collection activities. A significant shift in the ongoing pandemic is the fact that the type of secondary data needed and available is quite different from the classic incident reporting and coverage by local/national and international media as well as the international aid community. The currently available secondary data is to a large extent quantitative, with high volume and velocity. And on GitHub. While this may change once the pandemic hits countries with less developed information systems, the main sources of information will continue to be data on mobility from MNOs and data from health systems, followed by consumer and price data, and finally by employment and social welfare data. We will need to re-think how to use secondary data and what tools we need for its analysis. Understanding the information landscape is even more critical now, not only with regards to the availability of data but also in relation to documentation of approaches and data gaps, which will determine what our new baseline will be. The impact of the crisis on needs and standards will be measured just as with any other sudden-onset crisis: in the difference of conditions before and after the disaster. Furthermore, underscoring transparency of the data produced going forward, particularly with the constraints on hand, will be key to ensure confidence in its quality. National Statistics Offices and administrative data are already impacted and there will be gaps in time series data – which serve as baseline for major surveys and information systems, such as the ones for vaccine coverage and of course any population projections. Gaps in timeseries data such as labour statistics will pose significant challenges, as ILO points out: this may necessitate changes to the process of imputation and estimation used to generate estimates.

c. Act as Documentarian

We will have to act as documentarians with regards to our approaches, and even more importantly with regards to the challenges and gaps we face. Reducing bias, documenting bias, and being cognizant of who is being left out is key to the decision-making.

"Good documentation should make following the standard path of least resistance, rather than an uphill struggle for those designing data collection instruments allowing them to quickly identify the building blocks of standardisation they can draw upon"
Quote from Tim Davis' blog

Mapping and matching of open datasets, understanding their inter-operability, documenting approaches and linking up with non-traditional data holders will be a key element to support timely analysis and fill information gaps. What is more, the proper documentation of gaps and before/after

---

[5] https://github.community/t5/How-to-use-Git-and-GitHub/what-is-github/m-p/10069#M3417

comparisons not only of conditions but also of people's responses to our surveys will become a valuable dataset itself.

Reducing bias, documenting bias, and being cognizant of who is being left out is a key to the decision-making and thus an important element of the documentation needed.

Bias, Bias, Bias

We will have to think strategically and long-term in order to account for not only a lack of data from this current period but also for potential biases introduced by the shift in methods and budget in a time of overlap to allow us to understand those biases.

I will touch upon a few of the **types of biases that we can expect with the shifts described below**: coverage bias, non-response bias, response biases among which the most prominent are the social desirability bias, the extreme response bias, the demand characteristics bias. **Relying on data gathered solely through one app (and even one method) risks introducing statistical bias into our analyses of a particular issue**. Differences in individuals' access to the internet or to mobile phones could affect the way that different populations are represented, with conclusions based on this skewed data potentially reinforcing inequalities.

The Human Rights Data Analysis Group rightly highlights that the above described biases could also exacerbate blind spots. For example, incident reports collected through messaging-app data might give the impression that violence or a certain issue was concentrated in a particular part of a country (add particular socio-economic or ethnic group etc.), without taking account network outages or limited internet access.

Decisions taken during the design of interactions with messaging apps can also exacerbate biases. One instance would be to fail to account for the fact that in certain situations, women may not be able to access or submit data over a messaging app without the approval of a man in their household. For this reason, some interviewees described deliberate decisions to avoid messaging apps, instead using bespoke apps that included elements such as a requirement to sign in every time they used the service. A comprehensive breakdown of similar challenges and processes for mitigating them is beyond the scope of this article. For further information, see the Human Rights Data Analysis Group's outline of the key elements that can lead to inaccuracies when using data (including data collected in conflict-affected environments).

d. From Face-to-Face to Remote Technologies

We will have to shift from face-to-face to remote technologies, where mixed methods will remain key.

An obvious and relatively rapidly implementable shift will be the wider use of **phone surveys**. However, this comes with serious constraints beyond the loss of quality and the greater methodological complexity and biases, one of which being competition over resources and respondents. While it is true that phone surveys are cheaper than face-to-face surveys, they are still more expensive than web-based surveys as they require trained interviewers and a proper system for making the calls, both of which cost money.

Across those remote technologies, **there are four main issues that bring complexity to their use in social sciences: sample selection and attrition, measurement of irregular and changeable patterns, missing data, and researcher effects**.[6] As Sugie (2016) points out, those are particularly applicable to issues of concern with smartphone studies among hard-to-reach, highly mobile, and less technologically-skilled groups of people, which includes a large part of the population we work with in the humanitarian field.

There is a body of evidence studying the mode effects of the use of phone vs. face-to-face interviewing, as well as phone vs. cell phone vs. computer assisted internet interviewing (CAWI), but ultimately findings converge:

- For **sample-based probability surveys**, the mode effect is fairly consistently limited (meaning that phone interviews and face-to-face interviews provide comparable estimates for broad measures of population-based indicators, such as population-based disease and risk factor prevalence rates) and where it is persistent, it is based on social desirability bias.[7] [8]
- Contrary to some of the concerns expressed about a much higher non-response rate in remote data collection methods compared to face to face in guidance notes advising the humanitarian community on the use of remote technologies - the academic literature is finding the same or better response rates through computer assisted telephone interviewing (CATI) and CAWI approaches compared to face-to-face interviews with the caveat of the attrition effect in longitudinal studies.
- Regarding **phone vs. cell-phone effects**, academic research is fairly consistent here, too. As researchers have "caught up" with the number of cell phones they have allowed in their samples, bias has begun to decline. Research by Dutwin and Buskirk (2017), strongly suggests that the quality of data achieved through telephone samples is similarly "in the middle" of the two main alternative approaches, if not more closely aligned with in-person samples. Researchers interested in attaining the highest quality samples using the telephone would be well advised to conduct a majority of their interviews via cell phones.

---

[6] Sugie, N. F., (2016). Utilizing Smartphones to Study Disadvantaged and Hard-to-Reach Groups. Sociological Methods & Research, *47*(3), 458–491. https://doi.org/10.1177/0049124115626176
The provision of a new phone may change participant habits and routines for weeks or even months, depending on the novelty of the smartphone model, the tasks required for participation, and participant skill level.
[7] Mahfoud, Z., Ghandour, L., Ghandour, B., Mokdad, A. H., & Sibai, A. M. (2015). Cell Phone and Face-to-face Interview Responses in Population-based Surveys: How Do They Compare? Field Methods, 27(1), 39–54. https://doi.org/10.1177/1525822X14540084; Dutwin, D., Buskirk, D.T., (2017). Apples to Oranges or Gala versus Golden Delicious? Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples, Public Opinion Quarterly, 81(1), 213–239, https://doi.org/10.1093/poq/nfw061 ; Dutwin, D., Buskirk, T.D., (2019). Telephone Sample Surveys: Dearly Beloved or Nearly Departed? Trends in Survey Errors in the Era of Declining Response Rates, Journal of Survey Statistics and Methodology, smz044, https://doi.org/10.1093/jssam/smz044
; Dillman, A.D., Smyth, J.D., Christian, L.M., (2014). Internet, Phone, Mail and Mixed Mode Surveys: The Tailored Design Method (4th ed.). John Wiley & Sons Inc.; Frees, E. (2004). Longitudinal and Panel Data: Analysis and Applications in the Social Sciences. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511790928; Braekman, E., Berete, F., Charafeddine, R., Demarest, S., Drieskens, S., Gisle, L., et al. (2018). Measurement agreement of the self-administered questionnaire of the Belgian Health Interview Survey: Paper-and-pencil versus web-based mode, PLoS ONE 13(5): e0197434, doi: 10.1371/journal.pone.0197434. eCollection 2018; Sugie, N. F., (2016). Utilizing Smartphones to Study Disadvantaged and Hard-to-Reach Groups. Sociological Methods & Research, *47*(3), 458–491. https://doi.org/10.1177/0049124115626176
[8] While research of the prior decade finds that telephonic research has been quite resistant to bias stemming from non-response (Keeter et al., 2000; Groves, 2006; Keeter et al., 2006; Groves&Peytcheva, 2008), response rates for studies in that era were about twice as large as they are for similar studies conducted today. (Dutwin & Buskirk, 2019)

- **Web-based surveys** can produce higher quality data since an automatic skipping and branching logic as well as warning messages in case of missing or implausible answers can be foreseen. Web-based questionnaires, however, cannot be the sole mode of data collection for population surveys, as even in countries with high internet penetration, internet access and skills can vary across demographic groups. To overcome this limitation, a mixed-mode method in data collection could be used, including web-based, voice recognition, and/or paper-and-pencil modes, complemented by qualitative methods. (Thus mixed-mode within mixed-methods).
- Dutwin and Buskirk (2019) compared a battery of biases-related metrics for **non-probability panels, dual frame telephone samples, and high-quality in-person samples**. Unsurprisingly, results indicated that there is roughly a linear trend, with non-probability samples attaining the highest estimated bias, and the in-person sample the lowest. Results also indicated that bias estimates vary widely for the non-probability samples compared to either telephone or in-person samples.

In summary, if you must pivot to phone or computer-assisted surveys, do make sure that those are sample-based probability surveys. If you need to resort to non-probability panels, be very specific and thorough with the panel selection (see more below on panel selection). In 2019 Dutwin and Buskirk outline several methods of weighting of non-probability samples in order to minimise statistical biases significantly.

Most of the recent literature agrees that to mitigate bias and variance of coverage, mixed methods and two-frame (landline and cell phone) sampling are the best option, including in remote data collection settings. This requires however a considerable effort in the smart design of those tools and in ensuring alignment with said mixed methods. All in all, mixed mode surveys lower cost, improve timeliness, reduce coverage error, and improve response rates.

SURVEYWORLD.ME
Nous recherchons des participantes en Suisse, récompenses pour participation!

Most importantly for the settings of our work however that the above considerations are based on assumptions made in contexts of good phone and internet coverage. Access to phones in large parts of the world is not an issue. Yet, it is in some of the areas we operate in. Literacy and communication barriers (e.g. language) are another major constraint. While phone surveys are likely to be the most feasible mode in the coming year, we will need to outline and address the various barriers that are specific to the context in which we work in order to implement those.

Additional considerations will include questions such as: Do phone companies charge for incoming calls? What types of agreements with the phone companies are needed? If so, which companies cover which areas? If a phone is shared across multiple households, can it be used for multiple interviews? In the case of longitudinal studies, how do you ensure it is the same household responding each time?

A BRIEF NOTE ON SAMPLING, PANELS[9] AND LONGITUDINAL SURVEYS

Over the past weeks, you may have noticed increasing references to panel surveys. We could argue that sample-based probability surveys are often infeasible in a large number of the contexts we work in due to access constraints, the difficulty to find target populations, dynamic factors influencing the operational reality, and frankly the lack of financial resources. Carrying out repeated surveys of populations over time (or panel surveys) is something we rarely have the resources to do. It is, however, a method that is particularly useful to better understand dynamic relationships and to model differences among subjects over time.

However, if the interest lies in studying the time varying-economic, demographic, or sociological characteristics of an individual, then tracking individuals over time is much more informative than using a repeated cross-section. A panel survey design will yield more efficient estimators than a repeated cross-section design.[10] For example, it is only by tracking repeated observations on a sample of individuals that we can study the duration of a marriage, or time until divorce (a dynamic event of interest) or labour market outcomes (for which panel surveys were historically developed). In another example, The Consumer Price Index (CPI) is not considered a panel survey because individuals are not tracked over time. Such surveys are useful for understanding aggregate changes in a variable over time.

The difficulty in designing the sampling scheme so to reduce the risk of subjects leaving the study prior to its completion. Attrition can result in selection bias, which occurs when a rule other than simple random (or stratified) sampling is used to select observational units.

While the use of panel surveys as outlined above has grown significantly as a method of choice in internet surveys (in countries with high penetration of fast and reliable internet connection), they are not as straight forward to implement in contexts with the above described constraints by individual stakeholders.

Panels seem to work best on a large scale (i.e. nationally representative panel rather than one that is representative at the state or local level or of a specific population like students or IDPs) because the costs of establishing and operating (maintaining) them are substantial. Many surveys of both national and more localised areas have very specific population and eligibility requirements that are difficult to justify building an entire panel around. Some use panels because they can sample specific types of populations (e.g. college graduates, middle-aged men, etc.) but the ability to target other low-incidence and hard-to-reach populations is limited because in order to do so, the panel has to be quite large.

As such, only those who have the available resources can organise panels such as National Statistics Offices, Demographics Offices, large market research institutes, etc. Therefore, the UN Statistical Commission (UNSC) was one of the first actors to issue guidance to support the necessary shifts in approaches to the panel survey design during COVID-19. The UNSC suggested an adapted panel survey design in two phases: during and after lockdown. The new design requires that a panel survey in which every month's sample contains a subset of sampling units carried over from the previous month. In this design, it will be important, as also recently highlighted by the National Statistics Office of Colombia, to continue carrying out the overlapping sample for over a year and beyond the lockdown measure in order to allow measuring the mode effect.

To Phone Surveys and Beyond

While phone and web-based surveys will remain the most straight forward option currently available, there has been a significant investment in other technologies aiming to connect with the populations in need of assistance. SMS and Messenger micro-surveys and interactions with specific population groups can provide rapid verification of assumptions even on complex issues such as food security. Recent data suggests that worldwide, the average user spends four hours per day using their phone, 80% of which is spent on an average on just three apps (one or more of which are frequently messaging apps). The global ascension of messaging apps has been swift, surpassing SMS in daily message volume by 2013 and overtaking social networks in monthly active user accounts in 2015.

A very timely review by GOARN, IFRC, UNICEF and WHO on methods of community engagement through technology shows the breath of context-specific initiatives, ranging from Interactive Voice Response (IVR) surveys in Somalia to WhatsApp discussion groups, as well as everything in between. The caveat with those methods is that they are constrained by **unequal distribution and access to (smart) phones and coverage** across all groups, regardless of literacy (linked to both language and digital, gender and location, urban and rural).

This is a barrier to connecting with populations who have systematically lower access to information technology as well as education. If we are looking to conduct representative samples in an area with minority populations, it is quite possible that literacy levels can skew the response. Furthermore, there are contexts where minority groups are specifically banned from having phones. There are also situations where displaced or affected populations close to international borders are using the service providers of the neighbouring country (e.g. at the contact line in Ukraine) or use several SIM-cards accordingly – which makes it difficult to disambiguate and target with information.

As described, the trend of increased use of messaging apps is not uniform. Refugees living in urban areas are significantly more likely to own a smartphone compared to refugees living in rural areas. According to the most recent UNHCR overview, 68% of refugee households in urban locations had an internet-capable mobile phone in 2015, compared to just 22% in rural areas. Improved access for these groups can be facilitated by the distribution of (smart)phones for the purpose of the surveys (see the Vulnerability Analysis and Mapping – mVAM - project).

The cost of data mobile plans for internet access on a phone is another restricting factor. For instance, research by FireChat indicates that every month around one-third of smartphone users in Manila, Mumbai, and Mexico City do not use messaging apps over a certain period of time because they have spent all their mobile data allowance. However, the cost of mobile data has been forecast to fall worldwide in the coming years. Anecdotal evidence suggests that some groups, such as Syrian refugees, often consider paying for data as a top priority. Individuals also seem to find the overall cost of sending messages using a smartphone with a data plan to be lower than that of sending large numbers of SMS messages, even with a lower-cost feature phone. Additionally, some organisations provide free Wi-Fi internet hotspots or low-cost data plans in areas hosting high numbers of refugees.

---

[9] Frees, E. (2004). Longitudinal and Panel Data: Analysis and Applications in the Social Sciences. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511790928
[10] Ibid.

As UNHCR shows in a study as part of the Refugee Connectivity Project, effective use of connections requires skills, relevant content, and understanding of safety and security implications for refugees and marginalised communities alike. A connectivity challenge specific to refugees is the need pertains to providing identification to obtain a SIM card. Policies and practices should therefore account for this specific gap.

On another note, men report smartphone ownership at markedly higher rates than women in many countries. Interviewees working in crisis-affected areas added to this point emphasizing that this divide becomes wider among vulnerable groups in such areas. Others reported that even in situations in which women own or have access to phones, male members of their household often restrict or exert control over their phone usage.

> **i** **Examples from the Ebola Crisis in Sierra Leone, Cyclone Idai Response, UNICEF chatbots and natural language processing (NLP)**

SMS surveys: In the aftermath of Cyclone Idai in Mozambique, the use of the mobile surveying platform GeoPoll allowed to rapidly deploy an SMS survey to the hardest-hit areas inquiring about the current situation and needs. Surveys were free of charge and did not require participants to have mobile phone credit, enabling the participation of communities with lower socio-economic backgrounds. Surveys sent via SMS – which can be received by phones even when they are turned off or out of service at the time of sending – were able to collect quantifiable data informing the humanitarian response and needs before services are even deployed. Once the humanitarian response starts, mobile surveys can provide real-time feedback to aid organizations (read more on ANALP's Blog).

In 2016, the WFP Vulnerability Analysis and Mapping (mVAM) project conducted 250,000 surveys carried out in over 30 countries, asking nearly 4 million questions. Mobile technology offers a tremendous opportunity to improve communication in humanitarian settings. However, these new capabilities can also entail privacy and security risks for people and the communities where mobile surveys are implemented.

WFP has further proceeded to give people food price information for their areas with bots automatically reading the general WFP food price database (read the full WFP report).

Encrypted Chat Apps: Research suggests that bots could potentially increase the accuracy of survey data. This can be achieved by introducing questions that can only be answered by selecting one of several predetermined responses or using artificial intelligence to automatically recognize and highlight common data entry errors. In other words, the user is given tailored answers to questions that the bot provides by retrieving information relevant to the user from a database.

In 2016, UNICEF launched a chatbot integrated with U-Report, a service running in 28 countries allowing young people to answer polls and report on a broad range of development issues pertaining to their communities. UNICEF partnered directly with a team from Facebook to make the platform available on Messenger and also created integration for other apps such as Telegram. The bot is integrated with Facebook Messenger and Telegram, and UNICEF worked directly with a team from Messenger to implement the integration. It asks young people a series of questions on a weekly basis about issues that affect them, including education, sexual and reproductive health, access to health services, and their legal rights. UNICEF separately noted that Messenger has enabled them – and their partners who run the platform – to connect a wide range of countries. Users' answers are recorded in a database, analysed in real-time and shared in aggregated form on public websites and with decision-makers. When UNICEF and the U-Report partners receive unsolicited messages on a specific issue from members known as 'U-Reporters', UNICEF's partner organisations can log in and respond using a separate piece of software (CasePro) that recognises keywords relating to those partners' areas of expertise.

The above examples show that it already possible to use those technologies and that they can be effective in collecting the information we need. It is time that we stop just leaving drones, chatbots, and SMS surveys in the same "innovation corner" and start allowing them to take effect in our work. From the technical perspective, even a relative beginner in data science can build a chatbot in Python using libraries like scikit and NLTK. And the generation of these Data Science Information Management Officers has already arrived in the sector.

e. Focusing on Community Engagement

We will have to emphasise community engagement and capitalise on thick data.

This crisis has brought about not only an opportunity but the **vital necessity to directly engage with local communities in a meaningful way**. Indeed, detailed and regular research aiming at understanding local information ecosystems (e.g. language preferences or use of communication methods) is crucial. We need to systematically incorporate questions pertaining to access to and use of mobile phones into humanitarian needs assessments, preferred means and tools of communication, as well as language preferences. Phone surveys or SMS surveys, text apps are likely to be much more effective when tapping into local community structures. Furthermore, such approaches should also build on good practice such as the ones explored by Translators Without Borders, which asks the population if and how they understand assessment questions. When surveyed in October 2017, 63% of Rohingya refugees said that they were unable to communicate with humanitarian responders. The issue relates to communication barriers in a crisis scenario, which is exacerbated by the fact that three out of four persons within the refugee community are illiterate.

Focus groups, user testing, and surveys can help guide decisions on how people in a particular situation are likely to respond to the use of a particular communication channel. There are outright cell phone bans in certain conflict zones. Using mobile data collection in such settings would mean putting people in harm's way. Adequate assessment of the situation from a legal and policy perspective are therefore important to avoid this type of error. We should always keep in mind the risks of collecting data in conflict settings are simply too high. And when the risks outweigh the benefits, we are obliged to think of alternative approaches (e.g. remote sensing, social media monitoring). Thankfully, there is a wealth of work to build off from (e.g. the CDAC project, Internews, etc.).

The value of so called "thick data"[11]

With all the above described challenges of being removed from affected populations, of actual danger for the health of both affected populations and those who provide assistance, with limitations on the luxury to collect all of the nice-to-have information and store it just in case we need it, with the global impact to be expected of the COVID19 crisis that will result in severe disruption to lives and livelihoods, we have a much more complex task in understanding how different population groups cope. And we should expect these coping strategies as well as the disruptions that come along to be new. For example, the New York Times reported that response rates to phone surveys went up by 25% in the United States. The increase can be attributed to feelings of anxiety and concerns and reflects a need for mental support as well as the fact that people were at home during shelter-in-place orders.

---

[11] Loosely defined as data on people's emotions, stories, and models of their world.

Once again looking to the sociologists for advice may shed some light on this phenomenon. Social research is again urgently needed to document people's everyday experiences of living during the crisis, says Deborah Lupton, Prof. Professor and leader of the Vitalities Lab, University of New South Wales (UNSW) Sydney. Her work puts together an extraordinary collection of tools for field research during the pandemic which are all geared towards collecting not only behavioural information but also self-reporting information in real time, while participants are in their everyday environments. Complementing what we can observe on what people *are doing* real-time, with them self-reporting *how* they are doing will allow us to understand causal linkages and indirect impact of this global crisis in a much more holistic way.

From this viewpoint, we are uniquely placed to carry out research on temporary states or events that are often in flux. Real-time self-reporting is also well suited for hard-to-reach individuals, whose experiences are often more irregular compared to advantaged groups with stable routines. For these questions and populations, the ability to send surveys in real time improves upon interview methods, where error results from retrospective reporting and unnatural settings.

Oh wait, what about privacy?

I will not go into the complexity and risks that contact-tracing apps could introduce in the long-term. However, I do want to rewind and discuss privacy of individual and group level data in a little more detail. Most of the above-mentioned methods are evidently designed to collect individual level data and additionally include, or in fact produce themselves digital crumbs or digital trace data. It goes without saying that privacy and safety concerns begin even earlier in the process. For one, phone numbers must be lawfully obtained. It is important to avoid modalities that leave a large digital footprint which may expose people to risk. These include SMS and random digit dialling (RDD).

The more of this sensitive data we collect in the coming months, the more important it is to not only make sure our teams and operations are familiar with privacy standards and are operating responsibly. By the same token, we need to explore ways to share that data and make it available for further analysis to other stakeholders. This of course means that the utmost level of caution must be ensured and that includes the best technologies with regards to anonymisation, encryption and access control but also querying and analysis of that data.

Over the past year, we, at JIPS, have been privileged to have received support from the UNHCR Innovation Service to explore this aspect. And while there has been a healthy dose of scepticism as to whether the project is too ambitious or not, the COVID-19 pandemic shows us that we are actually somewhat late to try to build state-of-the-art solutions in research, academia and the private sector, in the humanitarian and development fields.

The more individual level data we generate and have in the coming months, the more prominent concerns will become that simple anonymization techniques are likely to be insufficient to ensure privacy is preserved (JIPS' report coming soon). The Imperial College London has already showcased this impressively, and the OCHA Centre for Humanitarian Data will carry out further research in what is referred to as the "Mosaic Effect" - which is even more pertinent in a context where we are capable to combine the data we have on what people are doing with the data on how they are doing and thus increase the risk of identifying individuals and groups that should be protected. The more we know, the more important is to make sure we not only use the data effectively but also ensure that it has the lowest possible risk of disclosure.