# BIG DATA FOR MEASURING THE INFORMATION SOCIETY: COUNTRY REPORT – THE PHILIPPINES

## CONTENTS


Contents ..... 1

List of tables and figures ..... 2

Tables ..... 2

Figures ..... 2

1. Background and context ..... 3

1.1. Project description ..... 3

1.2. Pilot country context ..... 4

2. Stakeholders in the project ..... 5

3. Getting access to the data: Procedures, legal documents and challenges ..... 6

3.1. Legal documents and challenges ..... 6

3.2. Resources required for the processing ..... 8

4. Methodology, technical description, tools ..... 8

4.1. Description of the data used in the project ..... 8

4.2. Reference data used in the project ..... 8

4.3. Access to the data ..... 9

4.4. Processing the data ..... 10

5. Recommendations ..... 12

6. Conclusions ..... 15

# LIST OF TABLES AND FIGURES

## TABLES

Table 1: Project timeline

Table 2: Reference data on the arrivals of tourists to the Philippines by month (April–May 2016)

Table 3: Quality assessment steps (verification list)

## FIGURES

Figure 1: Philippines 18 regions (LAU1 level and name)

Figure 2: Philippines 87 provinces (LAU2 level)

Figure 3: Scheme for processing the data in different tiers in the Philippines

# 1. BACKGROUND AND CONTEXT

## 1.1. PROJECT DESCRIPTION

The International Telecommunication Union (ITU) is the key official source for global information and communication technology (ICT) statistics. To complement ITU's current official data sources, methodologies and data sets used to produce these statistics and related benchmarks, ITU is pursuing innovative ways to utilize big data as a new data source. As part of this effort, ITU has carried out pilot studies in selected countries, including the Philippines, to produce new and policy-relevant ICT statistics with big data. The results of this project are expected to help countries and ITU to produce official ICT statistics and to develop new methodologies by combining new and existing data sources. This will directly benefit policy-makers and other ICT statistics users to benchmark and measure ICT trends in a compendious way.

Specifically, the project objectives were to:

- Combine the role of ITU as a standard-setting organization in terms of global ICT measurement with official data producers' experience and interest in working with big data: The ICT data producers included national statistical offices, regulatory authorities, telecommunication operators and Internet service providers (ISPs).
- Take advantage of ITU's involvement and experience in existing big data initiatives: ITU has organized a number of big data sessions in terms of regulation and monitoring, published information on the opportunities and challenges of big data, and developed a big data project on mobile phone data for health. ITU is also an active member of the United Nations Global Working Group on Big Data for Official Statistics, has networked with public and private organizations working on the topic, and has identified a number of potential (public and private) partners that would be interested and benefit from participating in this project.
- Explore and analyse the kind of new information society data and statistics that ITU and other stakeholders (including policy-makers, analysts, and other data producers and users) would value most, and which currently do not exist.
- Engage with ICT data producers to discuss the development of specific indicators for new statistics that could be produced through big data analytics, and by combining new and existing data sources and methodologies.

The first phase of the project involves several country pilot studies, which serve as concrete examples of how big data from the ICT industry (mobile network operators (MNOs) and ISPs) can be used and/or combined with existing official data (from ITU, national statistical offices and regulators) to produce new statistics, benchmarks and methodologies to measure the information society. This pilot study covers different types of big data sources, and several ITU regions. The pilot project used the methodology document developed by ITU (see annex), which includes the indicators and respective methodologies that were used for calculating the indicators in this project. The document includes 15 big data indicators, while the pilot country can also propose indicators that are relevant to the country's policy-making needs.

This report summarizes the pilot project conducted in the Philippines. More specifically, it highlights the roles, responsibilities and contributions of the different partners, the types of big data that were used, and the big data analytics that were carried out. It also provides recommendations on which and how data can be produced, collected, explored, aggregated, shared, analysed and presented. The report also

discusses data access agreements, and addresses issues and challenges related to possible data transfer, storage, privacy and confidentiality. It is expected to provide data producers with concrete steps on using big data to produce new or complement existing ICT statistics. An overall project report (forthcoming) will summarize the main findings from the pilot projects in six countries and make recommendations on how countries and ITU could use the results of the pilot studies for their ICT data collection. It will also make recommendations on how participation of private telecommunications and ISPs can be enhanced in future, similar projects.

## 1.2. PILOT COUNTRY CONTEXT

The Philippines is considered a very dynamic economy in the East Asia and Pacific region. Average economic growth of more than 6 per cent in recent years is mostly based on strong consumer demand accompanied by a young population and urbanization processes. The country is moving towards a leap from a lower-middle-income country to upper-middle-income status in the coming decades.

The Philippine ICT industry growth is consistent with gross domestic product growth and is expected to continue the positive trend supported by the financial, telecommunication, business process management and healthcare ICT sectors. The Philippine telecommunication industry is a major contributor to the country's economy. Continuous investments in communication equipment contribute to the growth of the ICT industry. With 121.1 million connections, the Philippines SIM card penetration rate is 119 per cent.

The Government of the Philippines continues its efforts to provide electronic services to its citizens and offer free Wi-Fi in more public places. The increased interest in developing Smart City solutions – with the focus on transparency and improving government efficiency by automating government services – also provides a good basis for growth in the ICT sector.

Two major telecommunication providers in the Philippines are Smart PLDT and Globe Telecom, which control the mobile market. In 2006, 3G service was launched in the Philippines, and 4G was made available in 2010, with an upgrade to Long-Term Evolution starting in 2012.

The Department of Information and Communications Technology (DICT) was established in 2016, and is the primary policy, planning, coordinating, implementing and administrative entity of the Executive Branch of the Government. DICT plans, develops and promotes the national ICT development agenda, and is the country focal point of the current big data project.

Initiated by ITU, DICT sees the important value of the project. With an active mobile phone market, DICT sees the potential of big data as a source of ICT indicators. While the country is still dependent on traditional statistics, DICT is cognizant of the potential of big data to support development planning and policy-making. Further to this, DICT also sees the importance of establishing public–private partnership models for data sharing and the development of new measurement standards for the ICT industry.

**Pilot project timeline**

The pilot project in the Philippines was implemented from May 2016 to October 2017 (Table 1). The project is still ongoing, as results from data providers are still to be transmitted.

**Table1: Project timeline**

| Activity | Time |
|---|---|
| Initial communication, stakeholder involvement and planning the project activities | May 2016 |
| Kick-off meeting of the project with all stakeholders in Manila | June 2016 |
| Preparations for processing the data by data providers, assessing the methodology for calculating the indicators and planning/allocating the resources for the tasks, resolving legal aspects | September 2016–April 2017 |
| Data scientist mission to Manila | 19–27 April 2017 |
| Calculation of the indicators (after the visit of the data scientist) | July–September 2017 |
| Assessing the resulting indicators, requesting some corrections from data providers, initiating the country report, getting feedback, comments and assessment from local stakeholders about the report | July–October 2017 |

The timeline of the project has been extended several times due to different administrative, legal and data-related issues. The initial plan for finishing the project in 2016 has been altered, due to the administrative and legal issues with data processing and data transmission, and has been postponed to late 2017.

## 2. STAKEHOLDERS IN THE PROJECT

The following local stakeholders were involved in the implementation of the pilot project:

- DICT is the executive department of the Philippine Government responsible for the planning, development and promotion of the country's ICT. DICT was the country focal point responsible for coordinating the activities related to the project in the Philippines;
- Data providers are:
    - Globe Telecom Inc. – One of the two biggest MNOs and ISPs in the Philippines;
    - Smart Communications – One of the two biggest MNOs and ISPs in the Philippines;
    - Philippine Statistics Authority (PSA), the national statistics office in the Philippines, which was responsible for providing reference data;
- The National Privacy Commission (NPC) is an independent body, attached to DICT for the purposes of policy coordination and monitoring, and ensuring the country's compliance with international standards set for data protection. It provided guidelines on the way to ensure confidentiality for personal and enterprises' information within the project.

DICT is interested in new indicators based on big data. DICT and other stakeholders see the potential of big data as a new data source to complement some of the existing ICT indicators, and to see if there are possible new indicators that can be calculated.

DICT also facilitated the coordination of activities related to the visit of the data scientist to the country, organizing the meetings with other stakeholders and even providing means of transport to move around the City of Manila. It must be said that all the staff from DICT involved in the pilot project were familiar

with the methodology and target outputs of the project, and were very helpful during the data scientist's visit.

## 3. GETTING ACCESS TO THE DATA: PROCEDURES, LEGAL DOCUMENTS AND CHALLENGES

### 3.1. LEGAL DOCUMENTS AND CHALLENGES

The information requested from data providers is confidential and subject to risk of disclosure (that is, individual information from data providers can be identified) and can be considered as sensitive for businesses. For this reason, the transmission and access to the data are usually regulated through legal documents, depending on the legal framework in the country. In particular, in the case of the Philippines, the legal documents required and signed between the stakeholders of the project were:

- an official letter from ITU to DICT for invitation to participate in the project;
- an official letter from DICT to ITU on its acceptance to participate in the project;
- an official letter from ITU to PSA to participate in the project;
- an official letter of invitation from DICT to PSA, Globe and Smart-PLDT (data providers);
- a memorandum of agreement between DICT and PSA;
- a memorandum of agreement between DICT and data providers;
- a privacy statement from ITU as requested by data providers;
- a DICT letter to NPC requesting for opinion on data protection and privacy requirements of the project;
- an NPC opinion on data protection and privacy requirements of the project.

Access to data was the main challenge encountered during the project, ensuring the confidentiality of the operators' data. NPC was consulted on potential privacy concerns, which resulted in a comprehensive NPC position paper on the possible confidentiality issues of the pilot project.

There was no clear idea from the beginning on how to manage the operational risks of the project, specifically on how to handle the confidentiality of personal information, and the risks of disclosure of enterprise information – sensitive information about the companies that provided the data. Mitigation measures were adopted by DICT in an ad hoc manner, as the requirements of the project became apparent. Legal templates were developed at country level in coordination with the legal departments of the data providers and the Legal Division of DICT.

Relative to data privacy, the position paper prepared by NPC included some requirements for the project, to wit:

- appointment of a Data Protection Officer;
- conduct of a Privacy Impact Assessment;
- adherence to data sharing requirements under WPC6/02;
- implementation of security measures.

Other mitigation measures adopted were:

- processing of data within data providers' premises to reduce the risk exposure of sensitive information;
- DICT assuming most of the risks in data confidentiality in the absence of a non-disclosure agreement between the data providers and ITU and data scientist.

The project also coincided with the change of administration after the 2016 elections, which saw a change of leadership in the then-ICT Office and the subsequent establishment of DICT. These political developments significantly affected the flow of documents within the organization to external partners.

The visit of the data scientist took place from 19 to 27 April 2017. Although it was previously agreed to have as many indicators as possible pre-calculated before the visit, in the end it was decided to start the visit without the accomplishment of this requirement, to speed up the pilot project in the Philippines.

To date, two of the three data providers were able to submit data by July 2017. The submission of the third is still pending, as its internal management has yet to give its full clearance to transfer the data.

The planning and allocation of resources were also challenges for data providers, with one data provider needing to hire a group of four to five people just for this project. Although the data providers said that they could summarize their data for most of the indicators, the linking of these data with the geographical data on administrative subdivisions of the Philippines (local administrative units (LAUs)) provided by PSA clearly added to the complexity of tasks.

During the visit of the ITU data scientist, it was agreed between the data providers that the indicators that could be computed were BD03, BD04, BD05, BD06, BD07, BD08, BD09, BD11 and BD13. As the data processing progressed, BD13 would later be dropped due to lack of data.

Finally, the main challenge is related to the confidentiality of the final indicators. This constitutes a major barrier to the dissemination of the results obtained. There are some special issues associated with the confidentiality protection of business data: businesses, especially large businesses, are more easily identifiable than households or persons, because the distribution of their characteristics is much more skewed. Thus, some measures must be taken to prevent the occurrence of a disclosure, measures that must be different from the ones usually taken to mask individual cases in household or personal data.

A general rule for business data is that figures are not published below a certain threshold of informants (three is generally the cut-off point).[1] Thus, the indicators computed for the Philippines require consent from the data providers to be published, at any level of breakdown or geographical disaggregation (whole country, LAU1 and LAU2 levels). Otherwise, there is a risk of disclosure of sensitive business information that can be considered strategic for their companies.

[1] See United Nations Department of Economic and Social Affairs, Statistics Division, *Handbook of Statistical Organization, Third Edition: The Operation and Organization of a Statistical Agency.* Series: F, No.88 (United Nations publication, Sales No. E.03.XVII.7), chap. XII, para. 547. United Nations, 2003.

One method to support research studies may be to obtain the consent for the disclosure from businesses prior to disseminating any data.[2]

### 3.2. RESOURCES REQUIRED FOR THE PROCESSING

An important challenge was the planning of the resources for the work of the data providers. Private companies usually cannot quickly re-allocate their resources to external, non-commercial projects (such as this project). As a consequence, the initial indicators had not been extracted and were not available to the ITU data scientist before and during the mission to the Philippines, as it was initially planned, but were calculated after the mission of the data scientist.

## 4. METHODOLOGY, TECHNICAL DESCRIPTION, TOOLS

### 4.1. DESCRIPTION OF THE DATA USED IN THE PROJECT

The project source data are composed of call detail records (CDRs) and Internet Protocol data records (IPDRs) gathered from data providers.

CDRs and IPDRs are well known and use mostly standard format, where each record represents an activity (i.e. call, SMS or Internet data usage session) performed by the subscriber, with attributes such as the time stamp of the beginning of the event, the antenna the device was connected to, the duration of the activity, and some other features, tracked for operational and/or billing purposes.

BD01 and BD02 indicators use antenna coverage maps for each technology (2G, 3G and Long-Term Evolution), which is usually provided in shapefile (SHP) format, a popular geospatial vector data format to be treated by geographic information systems. The data providers said that they did not have the geographical information about the location and coverage of the antennas and therefore BD01 and BD02 could not be computed.

The information to calculate the rest of the indicators (BD03–BD13) was provided by the data providers participating as stakeholders in the project. The data providers executed the matching of the location of CDRs and IPDRs with the geographical administrative information provided by PSA (different LAU levels) to produce the data requested for each indicator (including the geographical reference at the different LAU levels). The reference period of the data was the second quarter of 2016 (April–June).

### 4.2. REFERENCE DATA USED IN THE PROJECT

Reference data are required to be able to fully conduct the project according to the methodology. The reference data are used to link the data from data providers to geographical location (for geographical breakdown), and to calculate some specific indicators. As reference data, during the project, the following data sources (Figures 1 and 2, and Table 1) were required.

[2] See *Managing Statistical Confidentiality and Microdata Access: Principles and guidelines of good practice* (United Nations publication, Sales No. E.07.II.E.7), para. 67 and following, United Nations, 2007.

The first data source required was composed of subdivisions of administrative borders of the Philippines (LAUs), along with the urban–rural subdivisions, to be able to provide geographic and rural–urban breakdowns of the indicators. In the Philippines, apart from the whole country, two levels of local administrative units were considered: LAU1 – regions, and LAU2 – provinces (Figures 1 and 2). Data for the provinces were also split into urban and rural areas adding up the data of the corresponding barangays (LAU4 level administrative units), the level for which the distinction between rural and urban areas for statistical purposes is done. There are also some data for which the geographical reference is not available for data providers. In this case, it appears as "Unknown" to take into account the corresponding activities.

**Figure 1: Philippines 18 regions (LAU1 level and name)**

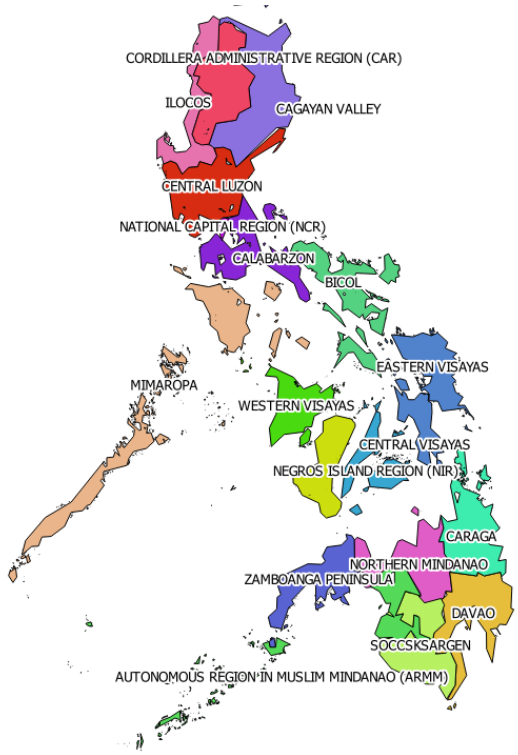



**Figure 2: Philippines 87 provinces (LAU2 level)**

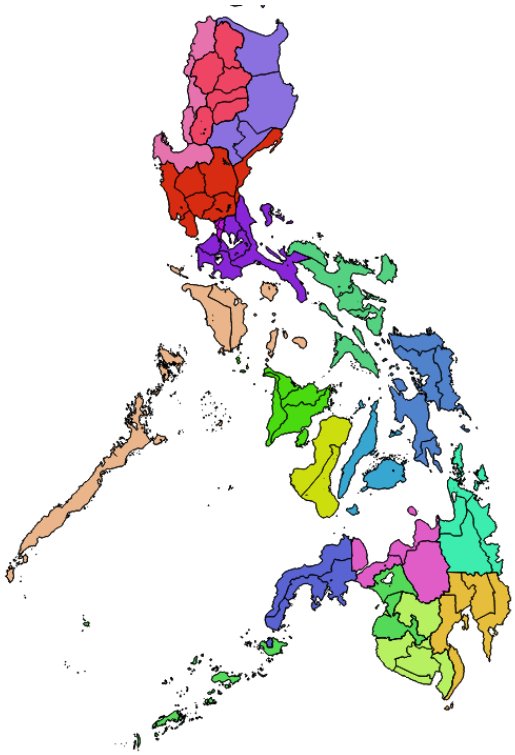


**Table 2: Reference data on the arrivals of tourists to the Philippines by month (April–May 2016)**

| | Foreign arrivals |
|---|---|
| **April 2016** | 471 598 |
| **May 2016** | 445 449 |
| **June 2016** | 459 138 |

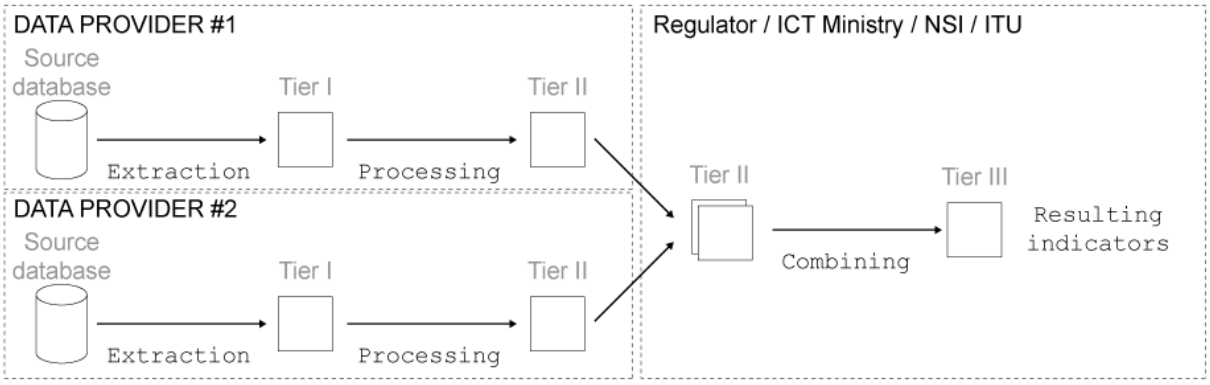
Source: PSA.

## 4.3. ACCESS TO THE DATA

For privacy protection reasons, and because initial raw data from data providers can be considered to include confidential business information, there are three tiers or phases of the data processing:

- Tier I – Initial, raw, not aggregated data extracted by the data providers from their databases and registries, that are the basis for calculation and which might include private and confidential business information. This can also be referred to as microdata.
- Tier II – Initially aggregated data per data providers with no private and some (or no) confidential business information, but which are still considered sensitive and kept from sharing with third parties.
- Tier III – The Tier II aggregated data from different data providers that represent the resulting statistical indicators of the project.

**Figure 3: Scheme for processing the data in different tiers in the Philippines**




For processing the data in the Philippines, the data providers proposed to process all the Tier I data inside their individual premises until Tier II, according to the method illustrated in Figure 3.

One of the challenges studied by NPC was the way to transfer the data to ITU and the data scientist, preserving the confidentiality of the information. Possible options analysed were:

- E-mail regular attachment: Apart from the risk of disclosure of data by anyone in e-mail copy, it could be a problem because attachments have frequent size limitations;
- Cloud, using one of the free or commercial providers of cloud services;
- Sharing of DICT account to upload and download files using Secure File Transfer Protocol.

The final decision, taken in coordination with the Legal Division of DICT, was to use the official DICT secure file sender service (pakete.gov.ph). The method allowed data providers and the ITU data scientist to upload/download files in a folder using a link provided by DICT that had a limited time (one week) of validity, allowing for DICT not to directly participate in the downloading of files and minimizing risk of exposure.

The Tier II files were downloaded by the ITU data scientist using this method from July 2017.

## 4.4. PROCESSING THE DATA

For the processing of the data, the data providers have IT departments that are competent for completing calculations (excepting minor issues), and some recalculation. From the initial processing, it

was possible to provide Tier II results to the ITU data scientist. Due to the confidentiality agreements, it is not possible to clarify which indicators were calculated by which data providers. It should be pointed out that the project is still ongoing for the Philippines, pending the submission of data.

Most of the reference data and the rest of the information from data providers were contained in Excel or comma-separated values files. These were obtained by the ITU data scientist through the file sender service Pakete.

Data submitted were subject to some quality and consistency checks and aggregated at other levels. The process made it necessary to come back very frequently and correct/modify some of the first data provided. The checks were based on previous information available at the general level for the whole country, and on some common sense and understanding of the indicators to produce (e.g. the traffic in a populated area cannot be "0").

The fact that the ITU data scientist received the aggregated data (Tier II) and not the detailed CDRs and IPDRs (Tier I), prevented performing verification and sophisticated tests to detect errors at the highest degree of detail. It has been assumed that the raw data met the requirements of not including duplicated records, empty records, records with NULL values, incorrect values and records outside of the pilot data observation period. Other checks and verifications could not be made that would confirm the validity of the algorithms against the methodology, so any potential known and unknown errors have not been explored. Such verification is necessary, although in the limited pilot project, and time-consuming.

The main issue with the lack of verification is that a thorough full-scale quality assessment could not be conducted, which means there is no 100 per cent assurance that the resulting indicators were produced exactly according to the methodology. The verification (quality assessment) steps are described in Table 3. Logically, if all verifications were conducted from top to bottom, the issues in the later phases should not exist.

**Table 3: Quality assessment steps (verification list)**

| Quality assessment | Possible errors | Comments |
|---|---|---|
| Sources and extraction of the microdata – Tier I, mapping the sources for different variables, compliance of variables to the proposed methodology, etc. | Not all necessary sources are included, extraction process excludes some variables, non-compliance of microdata variables to variables described in the methodology, etc. | Could not be performed. |
| Preparation and quality of Tier I microdata, naming convention and standard data format, logical consistency of the microdata, descriptive statistics of the microdata, etc. | Not all data requested exist, errors with geographical coordinates of the antennas, technology type of the antennas, etc. | Could not be performed. |

| | | |
|---|---|---|
| Processing of each indicator from Tier I into aggregated Tier II, algorithms (SQL, Python or other language) – how the processing of the data was conducted. | Non-compliance with the methodology, incorrect algorithms, incorrect aggregation (breakdown) logic, etc. | Could not be performed. |
| Tier II aggregated data structure compliance to the methodology. | Structure of the data set incorrect, naming conventions issues, aggregation "tree" and breakdowns incorrect. | Performed. Some errors occurred, so recalculation was necessary. |
| Logical consistency of Tier II aggregated indicators. | Indicators are not "logical", overcoverage or undercoverage issues. | Performed. |
| Combining Tier II data into Tier III. | Tier II data from different data providers cannot be combined. | Performed. |
| Tier III statistical disclosure control using suppression, rounding and interval publication. | Indicators with values under confidentiality threshold, indicators representing the status of 1-2 providers, etc. | Performed. |

After all the checks and refinements, the final version of each indicator's comma-separated values file was produced and all files were combined into Tier III resulting indicators.

## 5. RECOMMENDATIONS

Based on the experience of the pilot project in the Philippines, the following recommendations can be summarized for the Philippines, as well as for other countries:

**Administrative and operations**

- Identifying the necessary agreements and legal instruments required for conducting such a project, or continuous data request procedure, needs to be done well in advance of the commencement of the project (memoranda of understanding, non-disclosure agreements, official request letters, contracts, etc.). It would be very good to have standard legal and administrative procedures in place when the telecommunication regulator and/or national statistics office wants to request indicators based on big data in the future, although this might heavily depend on applicability, regulations and administrative relations of a specific country.
- The data protection authority's (or other privacy oversight organization's) involvement in the project is vital to make sure the processing of the data is conducted according to the legislation. It is also important that any necessary form of approval or recommendation is obtained from this organization at the start of any project involving private data.

- In this pilot project, the project coordinator and ITU data scientist were only able to visit the pilot country once each. This type of project would require multiple visits with intermediate calculations of the indicators, assessing the results and adjusting the methodology and algorithms.
- To obtain reliable results at the national level, inclusion of all informants (all MNOs and ISPs) in the country is necessary – otherwise, underrepresentation issues, technological and regional biases and confidentiality issues have to be dealt with, or undertaking of research projects on (non-trivial) methods for imputing and grossing up the data of non-informants' MNOs and ISPs is required.
- Agreeing on the scope of the publication of the resulting indicators in terms of confidentiality of the data providers is important, especially if the number of data providers is small – determining what is the threshold of confidentiality of the published data and how to ensure the safety of the business secrets of the data providers. It is important, in terms of public image, that local project organizers (focal points) position themselves as the guardians of the confidentiality of their informants (data providers). Thus, it is important to exercise great responsibility in the management of the risk of disclosure of confidential information.
- It is also important to have a strong coordinating mechanism to determine the roles and responsibilities for all stakeholders involved in the project, including ensuring that the responsible stakeholders have the appropriate resources, skills and time allocated for conducting the necessary tasks. Commitment of all stakeholders in an agreement should be signed to ensure the participation of the stakeholders in the project.
- It is recommended to obtain an initial assessment from data providers of the resources required and their availability to process the data within the agreed time-frame.
- It is beneficial to involve other potential stakeholders from the public and private sectors, either in the planning process or for presenting the results of the indicators. Additional indicators that can be related to ICT can also have big value for non-ICT sectors (e.g. origin–destination matrices, mobile money related indicators, etc.). Domains such as tourism, transportation, urban and regional planning, social sciences, etc., can largely benefit from this data source. Including universities in the projects can also boost the scientific research potential in a variety of domains and provide valuable feedback for data providers as well as public organizations.
- The promotion of projects including big data is recommended for good publicity and explanation to the public on how implementation of big data can help society.
- In order to avoid back and forth of data methodology, there is a need to have capacity building among the people involved with extracting of data.
- Funding avenues should be expounded for situations that may require complete system overhauls.
- There should be clearer and precise communication when making big data requests. Some of the explanations on what was needed may not have been very clear.

**Methodology**

- The specification, sources and business glossary and metadata of the raw data have to be specified and agreed between all data providers and stakeholders, including exact naming of the elements (e.g. what CDR represents exactly, what attributes are used in each data record, etc.).

- The methodology of the processing of the data for each indicator needs to be defined and confirmed by all stakeholders well in advance. The general understanding of how the indicators will be processed, what the intermediate tables are, how they should be aggregated, what the breakdown dimensions are, etc. – these questions need resolving before the actual processing takes place. This often requires a pilot project or iteration of methodology development before the final methodology can be agreed and locked for the "real" indicators.
- The proposed standards in naming and coding conventions should be followed by all data providers. Cleaning and reformatting the data by the receiving party can take a lot of time, and ideally this procedure has to be automated, so agreeing on standards of the data format and data exchange is important. Algorithms (even if they are prepared as examples) used to calculate the indicators have to be prepared along with the methodology, so no difference or misinterpretation of how the indicators should be calculated can appear. Again, this might require a pilot project for adjusting the methodology and correcting the algorithms several times, but the interpretation of the methodology by different data providers might cause serious discrepancies in resulting indicators.
- The reference data required for the data processing should be identified, obtained and formatted for the processing. The reference data include the geographical data about administrative subdivisions (multiple levels), population data (grid-based or administrative units), any existing indicators that are to be combined with big data indicators (e.g. BD13), etc. The lack of reference data, or using different types of reference data, can heavily affect the quality of the resulting indicators (e.g. if different data providers use different administrative levels, units or coding, the data cannot be combined).
- Validation methods of the steps of processing the data (extraction, processing and results) should be agreed between all stakeholders. How to validate the data extraction and processing of providers (and processors if they are different parties) is rather complicated. Here the validation of extracted data (distributions, descriptive statistics, etc.) and confirmation of exact algorithms used for processing can be the option. However, how to practically conduct that can be an issue (given that these might be sensitive business secrets of the data providers). The validation of results is often problematic if there is limited information on the technical algorithms used for calculating these indicators. It is important that national statistics offices take part in the validation of the indicators.
- It is important to understand that, in practice, the development of the methodology is a multiple iteration of preparing the methodology, creating algorithms, producing the results, assessing the results, adjusting the methodology, etc. This can be a time-consuming process.
- Agreeing on the source data, methodology, specific algorithms, business model and validation procedures takes a lot of time, so it is reasonable to plan an intensive preparation period including several meetings between stakeholders' business, legal and technical teams, to identify the possible risks involved and agree on the mitigation measures to avoid delays. During this project, the time for this preparation was very limited, but this can be considered a learning opportunity – after all, this is a pilot project designed to identify such issues.
- The value of the big data does not simply lie in the production of indicators with more breakdowns, but also provides the possibility to employ different data mining, machine learning, and other mathematical and statistical methods that were not practiced in this pilot project, due to limited time and resources. These more sophisticated methods can provide additional insights for ICT industry, but using such methods requires more intrusion in the data providers' databases and is therefore limited to external parties.

- It is also important to set the objectives of the development in ICT sectors that could be measured with such indicators. For example, the objective could be that, in the next two years, the number of subscribers with access to 3G and higher technology would be 75 per cent (BD05) for the whole country, and in no administrative subdivisions should this number be below 25 per cent. Or, the difference between urban and rural areas should not differ by more than 20 per cent. The specific target objectives are obviously country-specific and can be agreed between local stakeholders.
- System compatibility between systems used by data providers and those used by ITU should be established.
- Surveys should be flexible enough to allow data providers to issue data in the available format.

## 6. CONCLUSIONS

In the pilot project, a number of indicators related to ICT were proposed and calculated. There are two main possibilities to assess the value of these indicators:

- On the national level – If the indicators represent new insight into the ICT situation;
- On the international level – To compare the ICT levels of different countries.

As the country reports are generated after the results have been received, the comparison between the countries is presented in the Project Final Report. The comparison between the countries and the international value of the indicators can then be assessed. The international comparability can also provide additional insight on the value of the indicators within each country.

The pilot project in the Philippines has demonstrated the way in which big data can be used to produce indicators that are new or similar to existing ICT indicators. The results obtained enable ICT policy-makers and investors in the Philippines to make informed decisions based on evidence.

One of the objectives of the project was to test the process of generating new indicators based on big data and identifying the challenges and possibilities. This information is also valuable, as the challenges are often similar in different countries and the results serve as a good insight for other countries that wish to conduct such similar projects.

In general, the challenges can be divided into:

- administrative and legal; and
- technical and methodological.

Apart from the above-mentioned, one of the measures of the impact of the big data project in the Philippines may be on the lessons learned from the perspective of the proposed methodology. Some of these lessons are the following:

- Confidentiality issues can put in jeopardy the success of the project. As explained in section 3, if large businesses are included as informants, it may be difficult to make sure that outputs are

confidential.[3] For this reason, most of the indicators computed for the Philippines cannot be published at any level of breakdown or geographical disaggregation without the consent of data providers.

- Some of the magnitudes that the proposed indicators attempt to measure can be possibly better measured in other ways. One of the principles that statistical agencies should take care of when gathering data from businesses is that the way in which the information is collected must reflect the ways in which businesses keep records.[4] This pilot project may be the right time to learn on the issue.
- The ITU data scientist could have more access to detailed CDRs and IPDRs data for early detection of potential errors, and the ability to correct them, by exploring more detailed information and checking for outliers.
- This pilot project considers exclusively the information provided by some companies identified for the project and working in the country. For the national results to make sense and be comparable across countries, the rest of the companies must be taken into account, collecting or predicting their data.
- Some of the breakdowns proposed for the indicators may be not available for data providers. This does not mean that they could provide it in the future, if required.
- It is not conclusive if some of the indicators could be used to replace the existing indicator. It is clear that some of the indicators can complement existing ICT indicators, specifically for more information on geographical breakdowns, urban–rural distinction and the technological development. Monitoring the indicators over a longer period could help prioritize and focus the development of the technologies.
- This pilot project provided some of the indicators that can be calculated from big data in ICT, but further discussion on the usefulness, practical applicability and further development of those indicators is required. Based on the resulting indicators, it is expected that, in each country, some of the indicators can be important and used for national purposes, and some indicators are valuable for international comparability.

[3] See *Managing Statistical Confidentiality and Microdata Access: Principles and guidelines of good practice* (United Nations publication, Sales No. E.07.II.E.7), para. 67 and following, United Nations, 2007.

[4] See United Nations Department of Economic and Social Affairs, Statistics Division, *Handbook of Statistical Organization, Third Edition: The Operation and Organization of a Statistical Agency.* Series: F, No.88 (United Nations publication, Sales No. E.03.XVII.7), chap. XII, para. 527, United Nations, 2003.