The Data Revolution

There are numerous definitions of the data revolution. The report by the UN Secretary General's Independent Expert Advisory Group (IEAG) speaks of an "explosion" in the volume and production of data matched by a "growing demand for data from all parts of society" (IEAG, 2014). PARIS21 takes a complementary approach and refers to "delivering the right data to the right people in the right format at the right time" (PARIS21, 2015). This definition puts the emphasis on the fact that the data revolution should increase the use and impact of data on outcomes.

To enable this increase in use and impact of data, the strategies of National Statistical Systems where possible must include new data sources and increased engagement with new actors, such as the private sector, non-profits, and academia. These guidelines are written with a focus on this particular and important aspect of the Data Revolution. It is the access and use of these new data sources in a new data ecosystem of data users, owners, producers and legislators that will enable policy makers, civil society organisations and citizens to "monitor development progress, hold governments accountable and foster sustainable development" (IEAG, 2014).

The Data Revolution means different things depending on where you are in the data ecosystem. Official statistics and National Statistical Systems will face challenges in adapting to the new data environment. Models for statistical development which were implemented over the last 15-20 years may be bypassed by new data producing agencies and rendered irrelevant if countries do not adapt. The Data Revolution will affect every area of the National Statistical System. This is already happening in certain countries such as Senegal, where innovative approaches to planning and adapting statistical operations based on call detail records from mobile network operators already leverage new thinking into action. In other countries too, National Statistical Offices(NSOs) will need to adapt in order to maintain relevance in the new ecosystem.

The use of such new data sources (as defined later in this Section) is explicitly encouraged in the Fundamental Principles of Official Statistics. In particular, to honour citizens' entitlement to public information (based on quality, timeliness, cost), NSOs can draw on "all types of sources" (United Nations, 2014). The defining feature of official statistics is that they be provided by official statistical bodies according to professional standards and norms as laid out in the fundamental principles.

"Principle 5: Cost-effectiveness: Data for statistical purposes may be drawn from all types of sources [...]. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents." -- United Nations (2014), Fundamental Principles of Official Statistics.

All NSDS stages should respond to these new demands by considering:

- Further developing administrative data systems to produce reliable and robust population estimates to rebase population based data and anchor new data sources.
- Complementing traditional data collection with new data sources based on reviews

of cost, ease of collection, quality of data obtained through new processes and sustainability of the processes.

- Undertaking due process in evaluating cost effective substitution of existing data sources.

- Developing a comprehensive data plan and coordinated approach to data collection accounting for greater frequency in reporting up to nowcasting, greater disaggregation, more geographic relevance (see the Data Revolution Road Maps developed by the Global Partnership for Sustainable Development Data).

- Budgeting for staff/ human resources in the emerging field of data science, but also legal and regulatory capacity.

- Developing a plan to build new partnerships, either by building links with different actors within the private sector, tapping into the network of trusted data consultancies or leveraging regional statistical agencies to facilitate access to large multilaterals.

- Establishing strong links with Regional Strategies for the Development of Statistics (RSDSs) to combine regional resources in centres of knowledge and excellence where national statistical agencies don't have the capacity and resources to adapt.

- Reviewing existing statistical business processes and, if necessary, considering revision

Improving Existing Data Processes

There is no doubt that census and survey data will continue to be the key data source for international monitoring and national decision making and that systems for the collection of administrative data will need to be further developed. This legacy will continue and indeed, will always be relevant to rebase population based data. Reliable and robust population estimates will anchor new sources of data and render them more useful. The data revolution and its enabling technologies provide us with the tools to improve current data management mechanisms in several areas, such as:

1. **Survey methodology**. Improvements in survey methodology
2. **Mobile data collection**. Remote data entry using mobile phones and tablets
3. **Administrative data**. Current developments in administrative data availability and use (see in particular the chapter on Open Data)
4. **Data dissemination**. Coherent dissemination using new technology and innovative tools that adapt to user demand (see in particular the chapter on Data Dissemination)

Applying innovations can help streamline existing processes and assure greater reliability of survey data. NSOs must work with sectors and reconcile, promote and advocate for the development of strong administrative systems. These data are, in many cases, comparatively cheaper to use and though they are not population based, greater effort should be made to reconcile these data. In order to do so, the NSO should be collecting and centralizing key facility lists in order to help integrate the planning process. A clear goal that should be adopted by NSOs is developing a central geospatial reference area where

census boundaries, service points in health (clinics and dispensaries), schools (primary, secondary) and agricultural extension service points are all plotted together. This can serve as a strong reference data base for survey planning and stratification but also provide a service to civil society.

New Data Sources and National Statistical Systems

Mobilizing the data revolution for sustainable development requires NSSs to harness the exponentially increasing amounts of data, much of which is held by the private sector. New data partnerships with the private for-profit and non-profit sector can contribute to this by helping NSSs to save costs and provide more detailed and insightful data in a timelier manner, but they also come with several risks and challenges (cf. PARIS21, 2015). What is popularly called “Big Data” -- “traces of human actions picked up by digital devices” (Letouzé et al., 2013) -- will have to be managed and likely create necessary partnerships between academia, political analysts and the NSS. Care needs to be taken however since the use of these data still requires relatively sophisticated analytic techniques.

However new data may be defined, instead of attempting a definition of what constitutes a new data source, this chapter takes a more pragmatic route and narrows the scope to consider the following five data sources widely considered as “new” to official statistics, listed in order of feasibility for implementation in a developing country setting.

1. **Sensor and geospatial data**. Example: Using satellite imagery to estimate poverty levels (see here).
2. **Telecom data**. Example: Using call-detail records to estimate poverty and wealth (see here).
3. **Commercial transactions, including scanner data, credit card data, etc**. Example: Using scanner data for the Consumer Price Index (see here).
4. **Web crawling, scraping, search and analysis**. Example: Using online job board posting to estimate unemployment or LinkedIn data to estimate changes in job categories (see here).
5. **Social media**. Using Google Trends and sentiment analysis to measure subjective well-being (see here).

These data sources are particularly useful to report on indicators during inter-survey years and to capture changes in fast-moving indicators. Case studies from countries will be a primary source of information as use of these data is still highly embryonic. The NSS should play a role in developing greater understanding of country applications. But regional institutions will likely have to be actively involved to manage scarce resources and take advantage of economies of scale.

Access to Big Data (held by the private sector) and the related privacy issues are different from the use of administrative data (which is also sometimes referred to as Big Data). To access administrative data, NSOs can often rely on existing legal frameworks. Company data, however, are a new field and access modalities will need to be developed with national councils on privacy protection and all relevant stakeholders. At the international level, the UN Global Working Group on Big Data for Official Statistics is currently working on “Principles of Data Access” that can usefully extend the Fundamental Principles of Official Statistics (United Nations, 2014).

Robin, Klein and Jütting (2016) provide a detailed overview of the benefits and

complementarities as well as the risks and challenges associated with the use of new data sources for official statistics. The following points summarise the key lessons for NSSs.

Benefits and Complementarities

- **Cost effectiveness**. Public-private partnerships – defined as voluntary, collaborative agreements aimed at increasing an NSS' capacity to provide new or better statistics – can help NSSs save resources by both sharing data and avoiding high upfront costs in infrastructure for data management. First, the marginal costs of transferring data already collected by the private sector to an NSS stakeholder are extremely low. For instance, while a survey in the United States could cost over $20 million, matching private micro-data with existing aggregated data (e.g. linking plant level data to firm level data) could cost less than one-fifth of this amount (Landfeld, 2014). Second, by outsourcing the processing of the data, a capital-constrained NSS stakeholder can make use of the private sector's software and expertise, thereby avoiding high upfront costs.

- **Timeliness**. Since unprocessed mobile metadata is available quasi-instantaneously, Call Detail Records (CDRs) from mobile phone operators, for example, can yield near-real-time statistics.

- **Granularity**. Private sector data –CDRs and geospatial data in particular – can display great temporal, spatial, thematic and unit granularity. This is useful for the evaluation of short term policies, and the production of disaggregated statistics at regional and sub-regional levels, for example.

New data sources also enable statistical agencies to measure trends that were previously thought of as unmeasurable and to be more responsive to quickly changing policy requirements.

- **Data in new areas**. Big Data in particular has the potential of supporting the generation of new indicators, previously not compiled by NSOs, such as the measurement of inequalities which are especially relevant within the framework of the SDGs.
- **Increased responsiveness**. New data sources equip NSOs with the capacity to address new topics quickly and help academics to respond to what-if questions.

Risks & Challenges

Four challenges that relate to the particular properties of data distinguish most partnerships for statistics from partnerships in other sectors such as health or infrastructure: ensuring the security of proprietary data, creating a business model for data sharing, preserving privacy and coping with the technical difficulties associated with Big Data.

- **Access**. Proprietary information leakage is perceived as an important threat to for-profit and non-profit organisations. Data which provides actionable information about an organisation's clients, customers or strategy is most likely to be subject to secrecy. For example, CDRs, which are used by firms for geo-marketing purposes

are much more sensitive than public tweets, whichare relatively accessible. There are also concerns that governments might use the data for regulatory ends or that the release of data about an organisation's clientshurts their public image.

- **Incentives and sustainability**. Certain factors can reduce the attractiveness of data partnerships as a business model. First, uncertainty about the demand for big data can raise doubts about the extent of the market. Second, the benefits of data partnerships are not always immediate or straightforward. Third, there are concerns about the durability of new data sources. Indeed, given that private data is originally collected for non-statistical purposes, maintaining the extraction process can become a burden if the initial field of application loses relevance.
- **Privacy and ethics**.The data sharing dimension of data partnerships can jeopardize individual or group privacy. Thus, the security of personal and group information is both a condition for implementing data partnerships and a goal in itself. First, privacy legislation often imposes regulatory constraints. As most current privacy and data legislations do not specifically cover Big Data, existing laws are open to interpretation. Hence, NSOs do not have a clear mandate to exploit sensitive micro-data such as call detail records. Second, both public and private stakeholders face reputational and ethical issues: the simple fact that companies retain their customers' data can induce these to change providers. The transfer of data therefore poses an important risk to organisations.
- **Technical and statistical challenges**.These relate to the nature of most private data, Big Data in particular, which can often require specialised infrastructure and can be decentralized, unstandardized, unstructured and unrepresentative. The properties of Big Datadatasets therefore also impose restrictions on the structural characteristics of data partnerships, but also on the type of statistics that they can produce.

Integrating New Data Sources into an NSDS

The IAEG report on the Data Revolution called specifically to modify the NSDS approach to account for the Data Revolution by

> […] upgrading the "National Strategies for the Development of Statistics" (NSDS) to do better at coordinated and long-term planning, and in identifying sound investments and engaging non-official data producers in a cooperative effort to speed up the production, dissemination and use of data, strengthening civil society's capacity and resources to produce, use and disseminate data. – IEAG (2014, page 25)

The data revolution will change the way NSOs and NSSs operate and requires to get new actors involved in the NSDS process.

- **Changing Role of NSOs**:The changing data ecosystem of new data providers and users will result in changing business models for NSOs and other data producing agencies. Particularly, NSOs will be less vertically integrated and outsource more of their statistical processes. This comes with a change in NSOs' roles from ownership over the statistical production to ownership of the management challenges to assess risks and costs.

- **Changing skills profiles**.The changing role of NSOs also poses different requirements on NSOs skills set.NSO staff needs to have a proper command of new methodologies to identify, evaluate and access new data sources.This requires skills and training capacity in the emerging field of data science, but also legal and regulatory capacity.

- **Building regional centres of support**: Where national statistical agencies don't have the capacity and resources to adapt, RSDS should account for their coordinated approach in bringing the data revolution to national statistics. Areas where the data revolution could be leveraged to advance change at the regional level could be: (i) Providing centers of excellence and knowledge, (ii) Providing Big Data sandboxes: scalable and developmental platforms such as that of UNECE used to explore an organization's rich information sets through interaction and collaboration, (iii) Concentrating resources for key academic partnerships and promoting Public Private Partnerships to contribute to the pool of regional expertise.

- **Blended approach for compiling official statistics**: The degree of statistical generalizability of many non-traditional data sources is presently not well understood. Therefore, they should be employed with caution and traditional sources should be used to validate and calibrate these estimations, especially in the short term. Such a blended and complementary approach implies that NSOs will continue to rely on traditional statistical methods.

- **New forms of partnerships**: Access to new data sources requires new forms of partnerships.In recent years, we have seen the emergence of several successful cooperative structures, which often link different actors within the private sector. These can take time to build. Hence, NSOs must make the most of structures already in place. This can be done by tapping into the network of a "third-party" or by exploring less sensitive sources of data.There is also an important role to be played for closer cooperation between NSOs and regional statistical agencies.The latter can often facilitate access to large multilateral co-operations and reduce coordination costs. There is also scope for a closer co-operation between NSOs from developing and developed countries, for example through the sharing of satellite data.

- **Legal framework and protocols**. The success of data partnerships depends on the adoption of the systematic and transparent, protocol-based approaches to data sharing, which limit the risks of re-identifying individuals.Such protocols are already in place for sensitive medical data and essential in order to create trust in the reliability and integrity of national statistical systems when dealing with non-volunteered data.

- **Leading by example**. Different actors in NSS will take up different sources at different speed. NSOs will often be the lead agency in charge of formulating and implementing a country's NSDS. NSOs can play an important role by setting a good example for how new data sources for official statistics can be used by striving to experiment with new sources with due consideration of privacy and quality concerns.

New forms of partnerships

Access to new data sources requires new forms of partnerships. In recent years, we have seen the emergence of several successful cooperative structures, which often link different actors within the private sector. These public-private partnerships (PPPs) for statistics have three distinguishing features from PPPs in other sectors:

1. They need to be formulated as long-term agreements, as there is often a need for longitudinal data and, at the same time, few alternative suppliers exist – for

instance, phone logs are only held by a limited number of Mobile Network Operators.
2. Proprietary and privacy risks play a central role in PPPs for statistics, whereas in other sectors, for example, in infrastructure, risks are mainly linked to Value for Money and Return on Investment.
3. PPPs for statistics can cover any stage of the data value chain, including data collection, processing, analysis, dissemination.

To be solid, these partnerships can take time to build. Hence, NSOs must make the most of structures already in place. This can be done by tapping into the network of a "third-party" or by exploring less sensitive sources of data. There is also an important role to be played for closer cooperation between NSOs and regional statistical agencies. The latter can often facilitate access to large multilateral co-operations and reduce coordination costs. There is also scope for a closer co-operation between NSOs from developing and developed countries, for example through the sharing of satellite data.

Legal frameworks

An obstacle to obtaining and using new data sources is the need to define legal instruments and processes for accessing these data and preserve confidentiality of the user. These users are the private domain and form a rich source of client data that are guarded by the private interest that may hold the data. In many countries, the law is not clear on the legal processes for obtaining these data and placing them in the public domain. Much awareness was raised during the MDG era in placing survey microdata on-line for researchers. This entailed defining a protocol for documenting the data and defining standards. The result was the evolution of metadata definitions such as the Data Documentation Initiative (DDI) and the Statistical Data and Metadata eXchange (SDMX). In addition, a great deal of advocacy work was undertaken to change legal structures and re-define user rights accounting for anonymization processes. Instruments such as the African Charter for Statistics provided a legal backdrop for countries to guide their own legislative processes; The Charter indeed provides leverage and guidelines that help in modifying the law accounting for new data developments, such as the use of Big Data.

Tools:

The following is a list of concrete use cases of new tools for data management and new data sources:

The GWG Big Data Inventory is a catalogue of Big Data projects that are relevant for official statistics, SDG indicators and other statistics needed for decision-making on public policies, as well as for management and monitoring of public sector programs/projects.

The Advanced Data Planning Tool (ADAPT) is an innovative planning tool for statistical offices to adapt to new demands and changing data practices.ADAPT helps data producers in the national statistical system consult, cost and chart their indicators as defined by the national development plan. The tool is aimed at target countries trying to meet the demands of global agencies monitoring the SDGs and put these in context with their own national priorities.

The Platform for Innovations in Statistics (PISTA) is a collection of innovations in the area of data and official statistics in developing countries. It provides basic information such as reviews, contact details, brief assessments and case studies on institutional, organizational and technological innovations from both the public and the private sector.

The USAID Global Innovations Exchange is a global online marketplace for innovations, funding and resources in the domain of development. It aims to connect innovators with the

resources, contacts, and information they need to grow their innovation. The GPSDD Toolbox is a set of tools, methods and resources developed by practitioners from all regions in the world in the field of development data.

Good Practices:
A number of collaborative data sharing projects between the private and public sector have already emerged. According to Robin, Klein and Jutting (2016), these can be classified into four ideal types:

In-house production of statistics : Mobile network operator Telefónica has used its phone logs to develop several applications in-house, using its internal capacities. These projects demonstrate that private data producers not only are willing to help fill statistical gaps but also can derive benefits from using their data and resources for the public good.

Transfer of data sets to end users : The transfer of datasets to end user model consists of datasets being moved directly from the data owner to the end user. This model gives the end user more flexibility on what to do with the data. In this model, raw data is de-identified, sampled and aggregated to avoid possible re-identification. Between 2012 and 2015, Mobile Network Operator Orange organised two innovation challenges in which it made its anonymised CDRs available to research teams worldwide, despite the risks involved in terms of privacy and proprietary information.

Remote access: In the remote access model, data owners provide full access to their data to end users while still maintaining strict control on what information is extracted from its databases and data sets. Several examples of remote access exist already, such as the Data for Good initiative by Real Impact Analytics. In this project, the company accesses telecom data within the secured environment of the operators.

Transfer of data sets to a trusted third party : In the Trusted third party model, neither the data owner nor the data user support the security burden of hosting the data themselves. In the T3P model, both parties rely on a trusted third party to host the data and provide the necessary services to enable secured access to the data source. Since 2009, travel statistics for determining the balance of payments travel account are calculated based on call detail records thanks to a public-private partnership between analytics company Positium and the Central Bank of Estonia, Eesti Pank.

Moving algorithms rather than data: The model of shared algorithms allows the reuse of software by several private data owners who wish to perform similar analytical functions on one or several data sets. The Open Algorithms (OPAL) project aims to leverage the power of private data by providing an open platform and ready-made algorithms that allow private companies to run pre-defined algorithms autonomously in their own secure environments and only output only the aggregated results.

References:
IEAG (2014).A World that Counts: Mobilising the Data Revolution for sustainable development, Independent Expert Advisory Group on a Data Revolution for Sustainable Development

Landfeld, S. (2014). Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues. In: United Nations Statistics Division (UNSD) and National Bureau of Statistics of China, International Conference on Big Data for Official Statistics, Beijing, China: 8-30 October 2014

Letouzé et al. (2013), Big Data for Conflict Prevention: New Oil and Old Fires. In: Francesco

Mancini, ed., New Technology and the Prevention of Violence and Conflict, New York: International Peace Institute, April 2013.

PARIS21 (2015). A Road Map for a Country-led Data Revolution, PARIS21 Secretariat.

Robin, N., T. Klein and J. Jutting (2016). Public-Private Partnerships for Statistics. Lessons Learned, Future Steps:A focus on the use of non-official data sources for national statistics and public policy, OECD Development Co-operation Working Papers, No. 27, OECD Publishing, Paris.http://dx.doi.org/10.1787/5jm3nqp1g8wf-en

United Nations (2014), Fundamental Principles of Official Statistics, UN General Assembly Resolution 68/261, United Nations, http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf.