THE DEVELOPMENT ASSISTANCE COMMITTEE:
ENABLING EFFECTIVE DEVELOPMENT

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

Nicholas Robin, Thilo Klein and Johannes Jütting

OECD DEVELOPMENT CO-OPERATION WORKING PAPER 27

Authorised for publication by Brenda Killen, Deputy Director, Development Co-operation Directorate

OECD
BETTER POLICIES FOR BETTER LIVES

February 2016

Keywords: Public-private partnerships (PPP), non-official sources of data, national statistical offices (NSO), business model, big data

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

Nicholas Robin, Thilo Klein and Johannes Jütting[1]

ABSTRACT

Non-official sources of data, big data in particular, are currently attracting enormous interest in the world of official statistics. An impressive body of work focuses on how different types of big data (telecom data, social media, sensors and geospatial data, etc.) can be used to fill specific data gaps, especially with regard to the post-2015 agenda and the associated technology challenges. This paper focuses on different aspects of big data, but ones that are of crucial importance: what are the perspectives of the commercial operations and national statistical offices that respectively produce and might use this data; and which incentives, business models and protocols are needed to leverage non-official data sources within the official statistics community?

[1] The authors are thankful to Jos Berens, Christophe Demunter, Graham Eele, Gjergji Filipi, Christian Reimsbach-Kounatze and Michail Skaliotis for helpful discussions and comments.

The paper also benefitted from thoughtful comments from the following external reviewers: Jos Berens (Leiden University), Nicolas de Cordes and Stéphanie de Prevoisin (Orange), Siim Esko and Margus Tiru (Positium), and Till Zbiranski (Free University of Berlin). Comments were also gratefully received from Geoffrey Greenwell, El-Iza Mohamedou and Koffi Zougbede (PARIS21).

TABLE OF CONTENTS

EXECUTIVE SUMMARY: KEY MESSAGES 1

1. INTRODUCTION 2

2. CONCEPTUALISING PUBLIC-PRIVATE PARTNERSHIPS FOR STATISTICS 5

2.1. WHAT CONSTITUTES A PUBLIC-PRIVATE PARTNERSHIP FOR STATISTICS? 5
2.2. PUBLIC-PRIVATE PARTNERSHIPS FOR STATISTICS – OPPORTUNITIES AND CHALLENGES 6
2.3. EXISTING MODELS OF PUBLIC-PRIVATE PARTNERSHIPS FOR STATISTICS 8
2.4. PUBLIC-PRIVATE PARTNERSHIPS FOR STATISTICS IN DEVELOPING COUNTRIES 10

3. CASE STUDIES 11

3.1. THE IN-HOUSE MODEL – CENCELL, ALERTIMPACT AND SMART STEPS 11
3.2. A SYSTEMATIC APPROACH TOWARDS TRANSFERRING DATA SETS TO END USERS – ORANGE'S DATA FOR DEVELOPMENT CHALLENGES 13
3.3. TRANSFER OF DATASETS TO A TRUSTED THIRD PARTY– ESTIMATING INBOUND AND OUTBOUND TRAVEL IN ESTONIA FOR BALANCE OF PAYMENT STATISTICS 14
3.4. THE PREDICTION OF CONSUMER CONFIDENCE USING SOCIAL MEDIA DATA BY STATISTICS NETHERLANDS – A HYBRID MODEL 17

4. LESSONS LEARNED 19

4.1. THE PRIVATE SECTOR CAN DERIVE NON-FINANCIAL BENEFITS FROM SHARING ITS DATA 19
4.2. THE NEED FOR A STRUCTURED AND INCLUSIVE APPROACH 19
4.3. UNDERSTAND THE TRADE-OFF BETWEEN GRANULARITY AND LEVEL OF ACCESS TO THE DATA 20
4.4. FOLLOW A BLENDED APPROACH 20
4.5. CREATING A COST-EFFECTIVE MODEL 20

5. THE WAY FORWARD 22

5.1. SHORT-TERM – CONTROL SENSITIVE DATA 22
5.2. LONG-TERM – TRUST, EXCHANGE, SUSTAINABILITY 22

6. CONCLUSION 25

BIBLIOGRAPHY 26

ACRONYMS AND ABBREVIATIONS

CDR – Call Detail Record

D4D – Data for Development

DEEP – D4D External Ethics Panel

ICT – Information and Communication Technologies

MNO – Mobile Network Operator

NSO – National Statistical Office

NSS – National Statistical System

OECD – Organisation for Economic Co-operation and Development

PARIS21 – Partnership in Statistics for Development in the 21st Century

PPP – Public-Private Partnership

SDG – Sustainable Development Goal

EXECUTIVE SUMMARY: KEY MESSAGES

Public-private partnerships (PPPs) offer significant opportunities such as cost effectiveness, timeliness, granularity and scope for new indicators. There remain, however, a number of challenges, which need to be surmounted, such as technical difficulties, risks related to data confidentiality and a lack of incentives. A number of collaborative projects have already emerged and can be classified into four ideal types: namely the in-house production of statistics by the data provider, the transfer of private data sets to the end user, the transfer of private data sets to a trusted third party for processing and/or analysis, and the outsourcing of national statistical office functions.

Developing countries have a severe lack of resources and particular statistical needs. In this context it becomes important to harness the private sector's resources and use the most holistic models (in-house and third party) in which the private sector contributes to processing and analysing data.

We draw key lessons from four case studies: a) the in-house analysis of call detail records (CDRs) by mobile network operator (MNO) Telefónica; b) the production of travel statistics by trusted third party Positium; c) the Data for Development (D4D) challenge organized by MNO Orange; and d) the use of social media to predict consumer confidence by Statistics Netherlands. We derive the following key lessons from the case studies:

1. Private companies can derive non-financial benefits from sharing their data to obtain new analytical insights as well as more representative information about their populations of interest.
2. There is a need for a structured approach to PPPs that engages all stakeholders and in which rights and responsibilities are clearly outlined in enforceable contracts.
3. There is a trade-off between the granularity of data required and getting access to the data. This imposes a number of decisions on NSOs, depending on their priorities.
4. NSOs should follow a blended approach, combining unofficial data with traditional sources to reduce dependencies on the private sector and to "ground-truth" their results.
5. The pricing of data and analytics services should consider the financial means available to national statistical offices. Making use of intermediaries can cut costs through economies of scale in data processing.

1. INTRODUCTION

Progress in data collection and innovative applications of new databases have stimulated a growing demand for evidence-based decision-making. Most prominently, the need to monitor progress towards the post-2015 Sustainable Development Goals (SDGs) requires national statistical systems to provide more and better statistics. New data sources also have more targeted applications in areas such as urban planning (e.g. reducing traffic congestion), disaster relief and the management of pandemics. However, what has become a reality in many developed countries remains a challenge in most low- and middle-income states whose national statistical offices (NSOs) often lack the resources to generate the statistics necessary to guide national development plans.

Private actors have access to data, software and skillsets of which many NSOs could take advantage (IEAG, 2015, p. 3; World Economic Forum, 2015, p. 9). Public-private partnerships (PPPs) in statistics can help NSOs produce new indicators and improve current processes (e.g. increase the timeliness, cost effectiveness or granularity of official statistics) without being liable to important upfront costs. PPPs may involve outsourcing the collection or processing of traditional census or survey data but also leveraging new, unofficial sources of data. While surveys and censuses are frequently outsourced to private contractors, the recent proliferation of private sector data, big data in particular, has renewed the scope for PPPs in statistics and prompted calls for data-sharing, most prominently in the *A World That Counts* Report (IEAG, 2015)[2,3]. Unlike existing private-to-public administrative data streams, which governments require for regulatory or programmatic ends, PPPs for data-sharing involve the exchange of data for statistical purposes (see US OMB, 2014 for a discussion on administrative data). This paper takes PPPs in statistics as collaborative arrangements between the public and private sectors, which are aimed at increasing a national statistical system's (NSS) capacity to provide new or better statistics.

Recent years have seen a growing literature about the potential applications of unofficial data sources – especially big data – in official statistics and policy[4]. The United Nations' Global Sustainable Development Report (UNDESA, 2015, p. 144) has established an inventory of big data applications that could help monitor the SDGs. While there have been debates about the methodological viability of using such sources for statistics, there also remain substantial challenges related to incentives, ethics and regulation, which impede sustainable collaboration (see Lazer et al., 2014, on the

[2] In this paper, the terms: *private sector data* or *private data* refer to "non-official/unofficial" data (see footnote 3, for a definition of non-official data). This is not to say, however, that *all* data held by the private sector is unofficial data. For instance, a range of administrative data, which is used in official statistics, originates from the private sector.

[3] This paper defines *big data* as the "traces of human actions picked up by digital devices" (Letouzé et al., 2013). Human activity encompasses both actual behaviour (e.g. movements), and discourse (e.g. messages posted on social media). This definition encompasses certain business data sets (e.g. transaction records), but excludes a number of sources, such as online retailer prices, (which other definitions, such as those based on volume or variety, would qualify as big data), because they are not direct traces of human activity.

[4] In this paper, *non-official* data includes: a) private big data sources which are currently not used in official statistics (see PARIS21, 2012 and OECD), but also b) other non-statistical, unstructured or dispersed private sector data which are digitally collected (e.g. online retailer prices) but which are not direct traces of human activity and are not used in official statistics. On the other hand, *official data sources* are currently used in the production of official statistics. There is, of course, no clear distinction between both. For instance, CDRs are used in Estonia for official statistics, but in most countries, phone logs are considered non-official data sources.

limitations of Google Flu Trends). Recent events demonstrate that ad hoc co-operation is insufficient to reliably solve these issues. In 2014, for instance, negotiations for access to mobile phone data, which could have helped inform health authorities about the spread of Ebola, were severely hindered by their lack of structure and clearly defined roles (The Economist, 2014; World Economic Forum, 2015). PPPs, on the other hand, are characterised by agreements that clearly distribute roles and responsibilities, and address the aforementioned challenges in a systematic manner.

The present paper contributes to a body of literature that addresses the risks of and obstacles to co-operation between the private and public sectors in the field of data sharing and statistics. Existing studies include Ballivian and Hoffman's (2015) *Public-Private Partnerships for Data*, which draws lessons from past and existing multi-stakeholder agreements and offers a taxonomy of risks and benefits of data sharing.

The World Economic Forum's *Data-Driven Development: Pathways for Progress* report (2015) offers a comprehensive account of the challenges faced in harnessing privately held data for development as well as a number of solutions. In particular, the report stresses the reluctance of many private actors to share their data sets and proposes a set of actions to tackle different types of "perceived risks of sharing data" such as regulatory hurdles, the loss of proprietary information and threats to privacy (World Economic Forum, 2015, p. 10). The present study will draw on these issues, albeit through a case-study approach and with a focus on PPPs as a channel of co-operation.

Eurostat's *Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics* (2013; 2014a; 2014b; 2014c; 2014d; 2014e) also draws on case studies and offers an in-depth account of existing initiatives where call detail records (CDRs) were used for tourism and mobility statistics, as well as recommendations for setting up the technical and institutional infrastructure capable of producing official statistics from CDRs[5]. It is the most comprehensive study of CDRs as a source for official statistics: backed up by a large pool of case studies, it analyses the main obstacles to co-operation, namely strict or unclear regulations, privacy concerns, cost, corporate image and fear of proprietary information leakages. The present paper draws extensively upon the feasibility study but differs in its scope and approach. It concentrates on a broader range of data, which leads to several additional conclusions about how the properties of certain types of data can influence the structure of co-operation. Moreover, this paper is mainly aimed at developing countries.

Steve Landfeld's discussion paper on the *Uses of Big Data for Official Statistics* (2014) offers a focused discussion on the uses of big data in official statistics, with important sections on methodological issues and the protection of confidentiality. The study gives concrete and precise recommendations on data-exchange protocols. While Landfeld concentrates on data-sharing in the field of big data, this paper covers additional dimensions of PPPs such as outsourcing data processing.

This study will build on relevant examples to provide practical guidance for successful PPPs in developing countries. The case studies are made up of three examples in the field of mobile phone records and one in the field of social media. The important focus on CDRs was motivated by two factors. First, CDRs are the subset of unofficial data that has arguably generated the most interest by

[5] An MNO's *call detail records* are made up of metadata about its subscribers' calling and messaging activities (events). This typically includes but is not limited to the time of the event, the geographical coordinates of the closest antenna and an identifier of the person that the subscriber is communicating with.

public bodies[6]. Second, CDRs perfectly embody the conflicts of interests between individuals, corporations and NSOs that stand in the way of successful co-operation. The final case study highlights the existence of co-operation structures that can be built upon in the short and long term. Furthermore, while not all of the cases are PPPs, each has been chosen because it possesses characteristics that are conducive to co-operation between the public and private sectors.

This paper is structured as follows. Section 2 conceptualises PPPs for statistics, outlines the opportunities and challenges presented by PPPs, and presents existing and likely PPP models before discussing PPPs in developing countries. Section 3 examines the relevant case studies. This constitutes an important part of this report: precedents are crucial to build confidence among stakeholders and generate constructive collaboration, as suggested by the World Economic Forum (2015). Section 4 draws important lessons from the case studies. Finally, section 5 concludes with recommendations to generate public-private partnerships in both the medium and long term.

[6] Mobile network operators have been summoned during following several natural disasters and epidemic outbreaks. See ACAPS (2013), Call Detail Records: the Use of Mobile Phone Data to Track and Predict Population Displacement in Disasters, *ACAPS Reviews*, 12 June 2013. Orange's Data for Development challenge in Senegal was specifically aimed at development. Some countries already use call detail records for official tourism statistics. See Eurostat (2013, pp. 119-121).

2. CONCEPTUALISING PUBLIC-PRIVATE PARTNERSHIPS FOR STATISTICS

This section sets up an analytical framework for the case studies. It begins with a definition of PPPs, followed by an account of the opportunities and challenges associated with PPPs and an outline of the four main existing PPP models. Finally, the last paragraphs address a number of issues specific to developing countries.

2.1. What Constitutes a Public-Private Partnership for Statistics?

In this paper a PPP for statistics is defined as a voluntary collaborative agreement between the public and private sectors, which is aimed at increasing an NSS's capacity to provide new or better statistics. What distinguishes PPPs for statistics from other forms of public-private co-operation is the existence of an agreement which structures collaboration and defines roles, responsibilities and rights. PPPs are typically characterised as long-term agreements (see PPPIRC, 2015). This is an important requirement, especially for PPPs in data-sharing, where there is often a need for longitudinal data and where there are few alternative suppliers; for instance, in each country, phone logs are only held by a limited number of MNOs. However, outsourcing of a household survey can take the form of a one-off contract, as alternative suppliers are relatively common.

The public agencies of interest are mostly those that make up a country's NSS. These include the NSO, as well as other bodies in charge of producing official statistics, such as the central bank and various ministries. In the context of data, this paper will define the 'private' sector as the set of non-state bodies (corporations, non-governmental organisations, academia) that collect, transform and/or add value to data as part of their activities[7]. The private sector can conduct one or several of the following activities:

- Data collection, whether active (e.g. surveys) or passive (e.g. web scraping). For example, private firms sometimes conduct surveys or censuses on behalf of NSOs.
- Passive collection of data for the purpose of billing customers, targeting services and marketing (advertising, product suggestions, etc.). So-called 'data exhaust producers' profit from their own uses of the data and could possibly incur losses by sharing it, as emphasised below[8].
- Generate revenue from transforming or analysing raw data and reselling it.

The recent emergence of big data has resulted in the accumulation of valuable non-official data sets by the private sector, which has prompted calls for PPPs in data-sharing (see IEAG, 2015). However, the scope of PPPs is not limited to the exchange of data and can cover any stage of the data value chain, from collection to dissemination. Processing and analysis dimensions are particularly relevant to developing countries which often lack the resources to analyse large private data sets.

An important difference between traditional PPPs and PPPs in statistics is that the latter often involve handling sensitive information, which entails proprietary and privacy risks that are not as central in different sectors such as infrastructure, in which risks are mainly linked to value for money and return on investment.

[7] The private sector can also be divided into non-governmental organisations, academia and the for-profit companies, each of which can be viewed as a distinct subset of the private sector given that their fundamental purposes – and the incentives to which they respond – differ in many respects. This report, however, will not extensively focus on this distinction.

[8] This report has a broad understanding of "data-sharing" which encompasses the transfer of datasets, the granting of on-site access to databases, but also the exchange of statistics (as opposed to raw data).

The concept of public-private partnerships sometimes also refers to initiatives by public bodies to foster a data-driven economy by promoting greater use and exchange of data within the private sector or by disseminating its own data (Open Government Data)[9]. While this might lead to greater data generation and sharing in the very long-run, this type of co-operation is less relevant to developing countries and does not directly address the central issue at hand: enabling NSSs to quickly bridge the gap between the supply and demand of statistics. Finally, the statistical (re-)uses of administrative data initially transferred to the government for non-statistical purposes (e.g. regulation) does not qualify as a PPP, given that the data holders were under a legal obligation to share their data sets.

2.2. Public-Private Partnerships for Statistics – Opportunities and Challenges

PPPs have traditionally been praised for allowing a capital-constrained state to finance and deliver projects in a cost-effective way while distributing risk (Sabol and Puentes, 2014). In addition to these benefits, a relatively new data-sharing dimension of PPPs in statistics brings a new set of opportunities which are specific to data. This new dimension has also imposed a set of new challenges to PPPs.

2.2.1. Opportunities

Mobilising the data revolution for sustainable development requires NSOs to harness the exponentially increasing amounts of valuable and frequent data, much of which is held by the private sector. PPPs can contribute to this by helping NSOs to save costs, and provide more detailed and insightful data in a timelier manner.

- **Cost effectiveness.** Public-private partnerships can help NSOs save resources through both their data-sharing and value-adding dimensions. First, data is non rivalrous. Hence, the marginal costs of transferring data already collected by the private sector to an NSO are, in theory, extremely low (see Ballivian and Hoffman, 2015). For instance, according to Landfeld (2014) 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. Moreover, the cost of implementing a CDR processing infrastructure covering 10 million subscribers and with a 15 day latency period could cost little over 550 000 euros (Eurostat, 2014a, pp. 95-96)[10]. Second, by outsourcing the processing of the data, a capital constrained NSO can make use of the private sector's software and expertise rather than invest in a costly processing infrastructure.

- **Timeliness.** In theory, since unprocessed mobile metadata is available quasi-instantaneously, CDRs can yield near real-time statistics. In practice the degree of latency depends on the infrastructure and manpower available to process the data. Greater timeliness demands greater automation of the processing, which in turn has to be increasingly monitored. Nevertheless, a reasonable 15 day latency period can be achieved at a fraction of the cost of near real-time information (Eurostat 2014d, p. 20; Eurostat 2014e, p. 14)[11]. Monthly or bimonthly data remains attractive, given that the time lag between surveys and censuses often exceeds a year[12].

- **Granularity.** Private sector data – big data in particular – can display great temporal, spatial, thematic and unit granularity. First, even if it is difficult to provide the CDR based statistics in near real-time, they are based on near real-time measurements. Therefore, they can gauge the immediate effect of short-term policies *ex-post*. Second, mobile phone antennae are usually spaced several kilometres away from each other, but cell radii can fall below a thousand metres in densely populated areas. Therefore, mobile data offers high spatial resolution. In the case of tourism statistics, demand-survey data does not provide large

[9] For instance, the public-private partnership between Big Data Value Association and the European Commission aims at fuelling a data and innovation-driven economy in Europe. See European Commission (2014)

[10] To this should be added an estimated annual maintenance cost of 158 000 euros (Eurostat, 2014a).

[11] These comments are based on tourism statistics. The cost dynamics of timeliness, however, can be assumed to be common to all applications of CDRs.

[12] The Central Bank of Estonia's tourism statistics, which are compiled from CDRs, are available to policymakers on a monthly basis.

enough samples to determine the destinations of travellers. Mobile positioning data can break these down not only nationally, but also regionally (Eurostat, 2014d, pp. 36-37). Third, a number of characteristics can potentially be inferred from unofficial data, provided that they are subject to sound statistical analysis. Finally, big data is often micro data which, with regards to the unit of analysis, can be aggregated and disaggregated at will. Interestingly, a study by Blumenstock et al. (2015), which links individuals' mobile phone use history to their responses in a survey, has shown that CDRs can be used to model socio-economic characteristics at the level of the individual with reasonable accuracy. This suggests that the high resolution of big data (individual behavioural patterns, etc.) can, in some cases, be translated into equally granular information (individual socioeconomic information, etc.).

- **Data in new areas.** Big data in particular has the potential of generating new indicators, previously not compiled by NSOs, especially within the framework of the SDGs.

2.2.2. Challenges

Four challenges relating to the particular properties of data distinguish most PPPs for statistics from PPPs in other sectors such as health or infrastructure: insuring the security of proprietary data, preserving privacy, creating a business model for data-sharing and coping with the technical difficulties associated with non-official data sources.

- **Competitive risks.** Proprietary information leakage is perceived as an important threat to corporations. The degree of risk depends on the type of data and its relation to the company's business activity. Data which provides actionable information about a firm's customers or strategy is most likely to be subject to secrecy. CDRs, which can identify the location and behavioural patterns of an MNO's customers (and can be used for geo-marketing purposes), are much more sensitive than Twitter data (tweets being relatively accessible). There is also a concern in the private sector that their data might be used for regulatory ends (Landfeld, 2014).

- **Privacy and ethics.** The data-sharing dimension of PPPs can jeopardise individual or group privacy[13]. Thus, the security of personal and group information is both a condition for implementing PPPs and a goal in itself. First, both public and private stakeholders face reputational and ethical issues: as a study by Infas has shown, the simple fact that MNOs retain their customers' call data can induce these to change providers (Infas, 2010, as cited in Eurostat, 2014a, p. 100). The transfer of CDRs therefore poses an important risk to MNOs. Indeed, there have been cases of public opinion backlash in developed countries, where the media and civil society are active (ibid, pp. 102-105). In countries where these are not as strong, local reputation may not pose a significant threat to corporations. In this respect, international civil society and transnational organisations are crucial to insure that private information be handled ethically. Second, excessive violations of privacy can lead to decreased data quality and availability: fears about the uses and dissemination of their data can push people to provide less information or to change their measurable behavioural patterns, which could pose further challenges to statistical analysis (see Landfeld, 2014, p. 16). For instance, a study by forsa concluded that "as a result of data retention undertaken by MNOs in accordance with EU Directive 2006/24/EC (DRD), half of Germans would not contact marriage counsellors, psychotherapists or drug support services through telephone or e-mail" (Infras, 2010, as cited in Eurostat, 2014a., p. 100). Such changes in calling patterns significantly reduce the representativeness of CDRs.

- **Legal constraints.** As most privacy and data legislation does not specifically cover many unofficial data sources, existing laws are subject to interpretation (Eurostat, 2014e). Hence, NSOs do not have a clear mandate to exploit sensitive micro-data such as CDRs. In a survey conducted by Eurostat, regulatory and legislation barriers were the most recurrent obstacles cited by MNOs (Eurostat, 2014a, p. 202).

- **Incentives and sustainability**. Certain factors can reduce the attractiveness of PPPs as a business model. First, uncertainty about the demand for unofficial data can raise doubts about the extent of the market, especially since most NSOs have not worked with non-official sources in the past. Second, the benefits of PPPs are not always immediate or straightforward. Third, even if a PPP in data sharing can be beneficial to a company in the short term, it might involve costs in the long run. Indeed, given that private data is

[13] We would like to thank Jos Berens (Leiden University, Centre for Innovation) for an insightful discussion on the issue of group privacy. See also: Institute of Business Ethics and Orange (2015a).

originally collected for non-statistical purposes, maintaining the extraction process can become a burden if the initial field of application loses relevance (e.g. a hypothetical case in which MNOs no longer need to collect CDRs to bill their customers) (Landfeld, 2014).

- **Technical and statistical challenges**. These relate to the nature of most unofficial data sources, which can often be decentralised, unstandardised, unstructured and unrepresentative. The properties of these datasets therefore impose restrictions on the structural characteristics of public-private partnerships, but also on the type of statistics that they can produce.

Table 1. Summary of the opportunities and challenges associated with forming PPPs for statistics

Opportunities	
Cost-effectiveness	Non-rivalry of data
	Diffusion of fixed costs
Timeliness	
Granularity	Spatial granularity
	Temporal granularity
	Thematic granularity
	Unit granularity
Data in new areas	
Challenges	
Competitive risks	
Privacy and ethics	Reputational and ethical issues
	Decreased data availability
Legal constraints	
Turning PPPs for statistics into a viable business model	Uncertainty about the demand for unofficial data
	Demonstrating the benefits of PPPs
Technical and statistical challenges	

2.3. Existing Models of Public-Private Partnerships for Statistics

2.3.1. In-house production of statistics

In this type of arrangement, data is processed and analysed by the data provider, within its systems, which minimises any confidentiality risks. This can be understood as indirect data-sharing: the raw data is not exchanged, but relevant statistics are disclosed. Existing cases have involved a private company analysing their own data and publishing their methodology. This grey-box approach allows a certain level of transparency and would enable NSOs to verify whether the methodology complies with official statistical standards. A white-box variant, in which the NSO would have on-site access to the data and algorithms, would require a high level of trust from the private partner. A variant of this model would involve an intermediary processing and analysing the data from within the provider's servers.

There have been several cases where private firms have studied their data to address issues faced by the public sector. There is little information about whether any of these cases have featured a formal agreement, but these examples illustrate that the private sector has already operated according to this model of data-sharing. Examples include:

- the prediction of socioeconomic levels in Mexico by Telefónica (see 3.1)
- assessment of the effect of public health alerts on the spread of infectious diseases in Mexico by Telefónica (see 3.1).

2.3.2. Transfer of data sets to end users

Another sharing arrangement involves the physical transfer of databases to the NSO as the end user according to a sharing protocol with clear terms and conditions which specify the purpose of the agreement, the quality of the data, each party's responsibilities and the penalties for not respecting these (Landfeld, 2014). An advantage of this model is that NSOs can better review the results and more easily implement methodological changes. On the other hand, they rarely have the internal capacity to work with private sector data (e.g. big data), especially in developing countries. Furthermore, NSOs would face a range of legal and ethical concerns which are likely to be resolved in the medium- to long-term (see Section 5). Finally, companies currently take precautions which can limit the applicability and quality of the data (e.g. anonymisation, coarsening, or even modification of the data).

There is currently no example in which a private firm has agreed to transfer its data to an NSO on a long-term basis. Examples of corporations transferring their data include

- Data for Development (D4D) challenge launched by Orange (See 3.2)
- Analysis of a MNO's CDRs by the non-profit organisation Flowminder to study the spread of cholera in the wake of the 2010 earthquake in Haiti (see Bengtsson et al, 2015)

2.3.3. Transfer of data sets to a trusted third party

This model involves an intermediary analysing and disseminating (and possibly processing) data. A trusted third party is typically chosen for its ability to adhere to official statistical standards and its lack of incentive to misuse the data. Depending on context, this may be a private analytics firm, a regional statistical office, or any other competent and trusted body. This type of arrangement is particularly efficient when there is a need to combine data from several institutions: first, there is no need for the data holder to invest in data extraction; second, the data analysis will follow a harmonised process. Furthermore, in addition to promoting trust, the third-party approach can spare NSOs the upfront costs of a processing infrastructure. Here also, the grey-box approach is currently prevalent.

A PPP of this nature was concluded in Estonia, in 2009. Positium LBS extracts phone log data from MNOs, which it uses to calculate balance of payment statistics for the Central Bank of Estonia.

2.3.4. Outsourcing of NSO functions

According to this model, activities which are typically conducted by an NSO are outsourced to a contractor on grounds of efficiency. These can include traditional data collection, but also the processing and analysis of non-official data sources which are freely available (e.g. retail prices advertised on the internet). A distinction should be made between this model and the in-house and trusted third party approaches, in which the NSO does not have access to the data and is therefore contractually bound to outsourcing the processing and analysis.

PPPs of this type are widespread:

- National statistical offices have a long history of outsourcing survey and census data collection and processing. For instance, the Office for National Statistics in the United Kingdom outsourced several stages of the 2001 census of England and Wales, including the electronic processing of the census forms to Lockheed Martin.

- A more recent example of outsourcing is the Billion Prices Project, in which a daily price index is calculated through web scraping according to a methodology developed by MIT researchers. Several states make use of the index, which can be purchased from StateStreet (State Street Global Markets, 2012; Reimsbach-Kounatze, 2015).

2.4. Public-Private Partnerships for Statistics in Developing Countries

2.4.1. A lack of resources and bargaining power

As mentioned above, NSOs in developing countries face particularly tight budgets, which implies that they are both capital constrained and in need of cost-effective solutions. They rarely possess the resources or expertise to process raw unofficial datasets. In this respect, an in-house or a third party model would be a more feasible channel for data-sharing. Furthermore, PPPs would involve large transnational firms (e.g. MNOs) which have substantial bargaining power. These issues imply that regional statistical agencies such as AFRISTAT are likely to be more effective than individual NSOs at negotiating and implementing PPPs.

2.4.2. National Development Plans and statistical priorities

As outlined in the Informing a Data Revolution Roadmap, "a data revolution that works for developing countries is not just about generating more data. It is about providing the right data to the right people at the right time and in the right format" (PARIS21, 2015). PPPs should help fill the most important and relevant statistical gaps, as outlined by national strategies for the development of statistics (NSDSs) and guided by national development plans as well as SDGs. Given that the implementation of PPPs is likely to occur at a regional or international level, it is important that NSOs and supranational bodies actively engage with one another to determine national and regional priorities[14].

Achieving a system-wide approach within the NSS has been recognised as a key priority in developing countries which suffer from a lack of co-ordination between different data producers and between projects which are often funded on an ad hoc basis (PARIS21, 2014). The contractual nature of PPPs can give co-operative projects a long-term vision and anchor them within NSDSs. The third party approach would be most relevant in a situation where there is a need to aggregate data from several actors (e.g. different MNOs) through a harmonised and co-ordinated procedure.

Finally, developing countries are particularly affected by the "vicious cycle where inadequate resources restrain output and undermine the quality of statistics, while the poor quality of statistics leads to lower demand and hence fewer resources" (Badiee et al., 2004). By both engaging the private sector in statistical processes and helping NSOs provide more or better statistics, PPPs can help restore the private sector's confidence in official statistics and thus contribute towards breaking this vicious cycle.

[14] See Sanga (2011, pp. 113-114) for a discussion on the need for national-transnational co-ordination in setting focal points in the context of the MDGs.

3. CASE STUDIES

Each case study is an example of one of the first three aforementioned PPP business models. In terms of co-operation issues, the outsourcing model is less specific to statistics and akin to traditional PPPs. Therefore, it does not feature within the case studies. A fourth case study on the use of social media data is meant to illustrate how, for certain types of data, a market for sharing has developed and can be exploited by NSOs.

3.1. The In-house Model – CenCell, AlertImpact and Smart Steps

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.

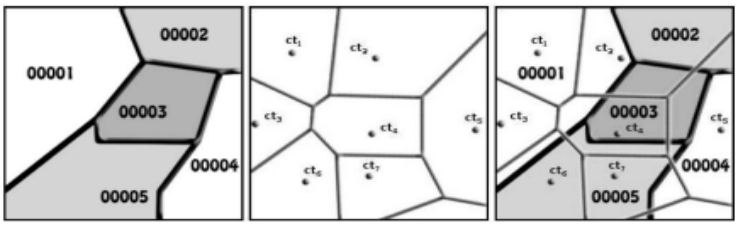
3.1.1. CenCell – low-cost and timely census approximation

The aim of the project was to explore the association between several behavioural attributes of phone users and their socio-economic level, with the aim of deriving a model capable of approximating a census at a lower cost and in a more timely fashion. Box 1 outlines the methodology followed.

Box 1. The methodology employed for bootstrapping CenCell

1) **Compilation of behavioural, social and mobility variables** based on subscribers' movements and cell phone usage (e.g. distance travelled every week, number of calls made every week, etc.).
2) **Determination of customers' place of residence,** based on a predictive model which assigned them a residential location, based on their calling patterns. The model was the outcome of a genetic algorithm which was applied on a subset of Telefónica's customers in a Latin American city – those which had a contract and whose home zip code was known. The genetic algorithm generated a number of candidate residential calling patterns (solutions) which were each assigned a fitness function according to their accuracy (the proportion of subscribers whose residential calling pattern was correctly predicted) and coverage (the proportion of subscribers for which calling activity was recorded during the proposed period). Candidates were successively filtered (between "generations") based on their fitness function until an optimal solution was reached. The optimal residential calling pattern was used to predict customers' home location. For more on the methodology employed for residential location, see Frias-Martinez et al. (2010).
3) **Matching the behavioural data with survey data**. Telefónica was able to obtain data about the socio-economic levels of 1 200 geographical regions (GR) in a Latin American city from the NSO. The proportions of each BTS that were overlapping with various GRs served as weights to calculate the socio-economic level (SEL) of each BTS ($SEL_{BTS}=_{p1}SEL_{GR1}+p_2SEL_{GR2}+\ldots+p_nSEL_{GRn}$ with each p_k representing the proportion of the BTS which overlaps with GR_k).
4) **Feature selection**. A subset of the 279 behavioural features was selected using the Maxrel and mRMR (respectively maximum relevance and minimum-redundancy-maximum-relevance) criteria, the latter being expressed either as a quotient or a difference between relevance and redundancy. Features selected according to maximum-relevance have the highest mutual information with the target class (here, the socio-economic level) (Peng et al, 2005).
5) **Classification**. Support Vector Machines are used to test which feature combinations best predict the SELs. The data set (made up of vectors of 1 to 279 dimensions, depending on the number of features chosen) was divided into a training and a test set. For each subset of ordered features, the training set was used to determine the optimal, accuracy-maximising values of an RBF Kernel's parameters (the Kernel function is used to transform the vectors into a dimension in which they are separable by a hyperplane, i.e. classifiable). The test set was used to test the SVM model associated with each subset of features. Accuracy does not substantively vary beyond 50 features. Thus, it is possible to focus on the most accurate model containing less than 50 features. The two best models contained the top 38 and 17 features (each respectively 80% and 79.1% accurate), determined according to maximum relevance (Maxrel).
6) **Another classification method, random forest,** displayed classification rates of up to 82.4%, when 38 features were selected. The maximum classification rate using Support Vector Regression Machines, at 80.13%, was slightly lower than the one achieved using random forest.

Figure 1. Stylistic illustration of the overlap between Geographical Regions and the Voronoi polygons approximating BTS coverage areas



(a) Geographical Regions (GRs) which are assigned a socio-economic level

(b) Approximation of the coverage areas of BTS

(c) Overlap

Sources: Soto et al. (2011) and Frias-Martinez (2011).

3.1.2. AlertImpact

Another application, AlertImpact was developed in response to the H1N1 flu epidemic in Mexico as tool to assess the effectiveness of health alerts in curbing the spread of epidemics (Frias-Martinez, 2012). It used CDRs to estimate actual population mobility during the alert, which was compared to a hypothetical model in which no alert was given. The reduction in infection rate attributable to policy action was inferred from the difference in mobility (Frias-Martinez, 2012). AlertImpact's algorithms could potentially be combined with Smart Steps, an application which also analyses mobility patterns, therefore offering scope for joint processes.

3.1.3. Smart Steps

In partnership with the market research firm GfK, Telefónica Dynamic Insights designed a tool capable of measuring the size, characteristics (demographics, residential area) and behaviour of crowds using anonymised and aggregated mobile phone data (Stone, 2015; Telefónica 2012)[15] [16]. This application was successful among retailers and transport companies. For instance, it allowed Morrisons, one of Britain's main food retailers, to launch a successful coupon advertising campaign which led to a "150% increase in new or reactivated customers visiting Morrisons without any reduction in average customer spend" (Telefónica and GfK, as cited in Stone, 2015). The example of Smart Steps demonstrates how a mobile network operator was able to add value to its phone logs through providing a service to third parties, i.e. external monetisation, rather than only using them for their own marketing campaigns (Stone, 2015). Moreover, Telefónica has expressed an interest in working with the public sector, both through Smart Steps and other pilot projects such as the prediction of human behaviour during floods (Telefónica, 2012; Pastor-Escuredo, D. et al, 2014).

Furthermore, the commercial success of Smart Steps suggests that it could be applied to the field of official statistics at a low cost. Indeed, Telefónica can recover the fixed costs of creating the application through its market of private clients. Thus, it could afford to charge NSOs a fair price for its statistics (equal to their marginal costs), provided these draw on the software developed for Smart Steps (i.e. if both follow joint procedures).

3.2. A Systematic Approach Towards Transferring Data Sets to End Users – Orange's Data for Development Challenges

Between 2012 and 2015, Mobile Network Operator Orange organised two innovation challenges in which it made its CDRs available to research teams worldwide, despite the risks involved[17].

3.2.1. The data

During the Senegal Challenge, three CDR based data sets were provided by Orange (de Montjoye et al., 2014; Blondel et al., 2013):

- hourly antenna-to-antenna traffic: the total duration and number of calls between antennae

[15] Smart Steps is part of Telefónica Dynamic Insights, a business unit geared towards monetising big data (Stone, 2015).
[16] Although Telefónica has not explicitly mentioned the type of mobile phone data that Smart Steps was based on, CDRs (as opposed to active mobile positioning data) are the most likely source. Furthermore, both are sensitive and associated with the same risks (Eurostat, 2013, p. 186).
[17] These were mostly related to privacy and proprietary information. The latter was particularly significant in Côte d'Ivoire, where Orange, with a market share of one third, faced competition (SciDev.Net, 2013). Legal risks were mitigated, as the D4D challenge took place under the oversight of the "Commission de Protection des Données Personnelles".

- individual trajectories (antenna level/two week period): this data set provided information about the trajectories of randomly sampled individuals at a high resolution, but only over two weeks
- individual trajectories (arrondissement level/entire observation period): this data set contained the trajectories of the sampled individuals over the entire period but was coarsened to hinder re-identification.

Antennae locations were blurred, as Orange considers their exact position to be sensitive information (Blondel et al. 2013). The data were subject to enhanced anonymisation by the Orange Group and its local subsidiary (Orange, 2014).

3.2.2. A systematic and ethical procedure for sharing CDRs and assessing research projects

In order to access Orange's CDRs, research teams were screened and required to agree to terms and conditions (Institute for Business Ethics and Orange, 2015). The papers produced by researchers were assessed by a review panel according to a specified framework which was based on methodological and ethical assessments of the projects. The whole process was governed by a clear structure which comprised two bodies in charge of ethics. While one was internal to Orange, the D4D External Ethics Panel (DEEP), was made up of representatives from a range of external entities, and was in charge of advising the internal workgroup. Interestingly, the panel was not confined to assessing the research ethics, but also examined the implications of the projects. For instance, one paper which touched upon a sensitive political topic, was selectively presented to relevant authorities rather than publically released, in order to prevent social unrest (Institute for Business Ethics and Orange, 2015).

3.2.3. Building on the D4D challenge: Open Algorithms for NSO (OANSO)

Building on the success of the D4D challenge, Orange is currently developing a business model for sharing its data with NSOs. A variant of the in-house approach, it is based on open software, which likens it to a white-box model. Algorithms would therefore benefit from both the private sector's experience in data science and NSOs' expertise in official statistics methodology. Following a process of certification by trusted bodies (e.g. national or regional statistical offices), the algorithms would operate behind the MNO's firewall (de Cordes, 2015). This initiative once again demonstrates that private data holders see potential in PPPs.

3.3. *Transfer of Datasets to a Trusted Third Party– Estimating Inbound and Outbound Travel in Estonia for Balance of Payment Statistics*

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.

3.3.1. The Positium Data Mediator (PDM)

Positium is independent from both Eesti Pank and telecom providers and has more than ten years' experience working with MNOs to produce mobility statistics, therefore acting as a trusted third party between the Estonian national statistical system and telecom providers (Kroon, 2012). The partnership between Positium and Estonian MNOs addresses telecom operators' most important concerns regarding CDRs, namely the preservation of business secrets, the protection of their subscribers' privacy and the compliance of processes with privacy legislation. For instance, EMT (Estonia's largest MNO) gives each tourist a unique pseudonymous ID number through a one-way algorithm and

the representation of data follows a procedure that further reduces the risk of re-identification (Kuusik et al. 2010). Positium's methodology also has the approval of the Estonian state data protection agency (Kuusik et al, 2010). Moreover, through its secure servers, Positium possesses the technical capabilities to safely handle the data provided by MNOs (see Ahas et al., 2007 and Kuusik et al. 2010). Nevertheless, Positium continues to encounter issues with data access which is still limited (see Ahas et al. 2007).

3.3.2. Methodological issues

Box 2 describes the methodology for calculating inbound and outbound travel statistics. The fourth paragraph, on estimation, illustrates Positium's blended approach to using non-official data, by combining its estimates with traditional sources for purposes of calibration and interpretation.

Box 2. Positium's methodology summarised (inbound and outbound tourism statistics for the calculation of the balance of payment's travel item)

1) **Event extraction – which includes the sending and receipt of calls and text messages – from the MNO's log files**. A given MNO's call detail records should make available the identity of the subscriber linked to the event, the time and duration of the event, the identity of the antenna where the event was recorded, the country code of the subscriber (for inbound travel) or of the country in which the event took place (obtained through the code of the MNO which provides the roaming services). The user's country of residence can be determined according to the SIM card's registration country. This assumption, however, does not always hold. For instance, it is possible to purchase international SIM cards in Canada, regardless of one's residency. Therefore, owners of Canadian international SIM cards should be filtered out as most likely non-Canadian.

2) **Initial processing**. Non-representative and black-listed (e.g. for security reasons) data is removed, and the data is formatted for further handling. The data is also filtered in order to control for quality and eliminate errors (missing data, duplicated data, missing attributes, etc.) The pseudonymisation of subscribers takes place before further analysis.

3) **Sample frame formation**. Positium has established inbound and outbound travel identification algorithms to calculate the travel item of the balance of payments. These determine the duration of the stay and eliminate events which are not considered as travel. In the case of inbound roaming, local MNOs are not informed when international roamers return home. Therefore, travel patterns should be analysed in order to distinguish pairs of events which belong to a same visit from those which belong to different travels. According to statistics, events which are separated by more than seven days are more likely to be part of distinct visits and are therefore considered as such. In the case of outbound roaming, the domestic MNO has access to both local and "foreign" events. It is therefore easier to determine the end of an international visit. However, in the absence of any call activity, it should be assumed that, after six days, the traveller has returned. In both cases "cross-border noise" (calling activity taking place near a border and within the network of a foreign operator), transit travel (of short duration and located within major transportation links such as airports, ports and important roads) and long-term stays are eliminated from the frame.

4) **Estimation**. The resulting estimations are compared to and corrected with the help of various reference statistics which include accommodation statistics, ferry statistics and foreign surveys. These figures should be the lower bound of the mobile positioning results.

Sources: Kroon (2012), Eurostat (2014a), Eurostat (2014b), Ahas et al. (2014)

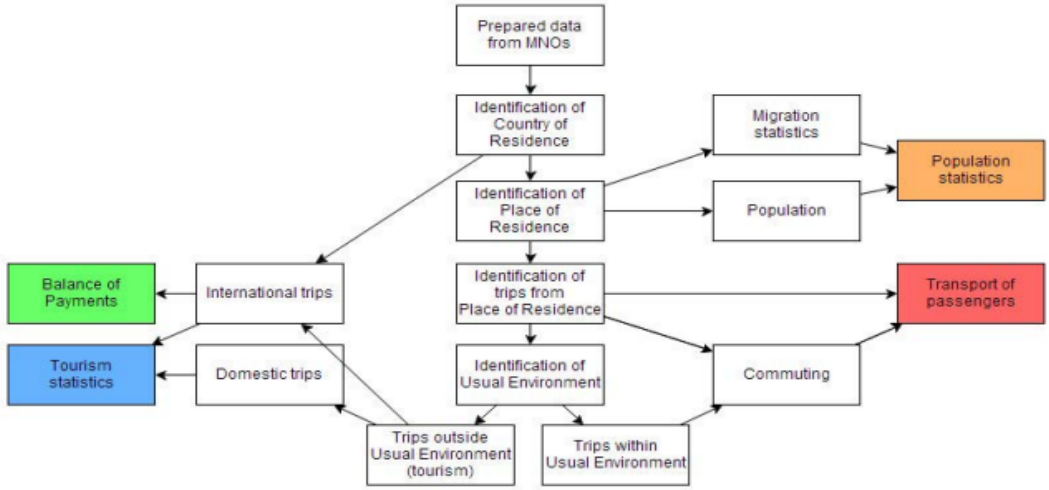
There are two ways to combine the data from different MNOs (Eurostat, 2014b, pp. 57-60). Either the micro-data from all the MNOs can be pooled together and subsequently aggregated, or these can be separately aggregated before being combined. While the latter method is vulnerable to various caveats such as cross-roaming – travellers that change operators while visiting would be counted more than once (the different MNOs would have no way of knowing which of their subscribers has also used their competitors' networks), it is more secure, as MNOs would only share aggregated results. Concerned with the privacy of their subscribers and the security of their proprietary information, Estonian mobile network carriers currently each employ their own anonymisation hash function, which implies that the Positium Data Mediator has to separately aggregate the data from each MNO. In order to correct for the over-estimation of visits inherent in separate aggregation, Positium uses existing data about cross-roaming to determine correction coefficients: these will reduce the amount of visitors consistent with the average amount of times that they change operators.

A further issue is that different MNOs are unlikely to have identical penetration rates among various strands of the population. Positium accounts for this by assigning weights to each MNO's estimates for a specific population strand. Weights are determined according to the penetration rate of the MNO among the given population segment and the type of contract between the MNOs in their country of origin and the local MNO. If both operators have a close connection, the weight assigned to the given population will be higher.

3.3.3. An attractive business case

The partnership with Eesti Pank offers Positium as well as Estonian MNOs important non-financial benefits. First, the software and methodology that Positium developed for providing Eesti Pank with international travel statistics has been used in other projects such as a study of international commuting (Eurostat, 2013). Indeed, as shown in Figure 2, there is scope for joint processes in the field of mobility statistics, as certain methodologies can be used for several purposes (Eurostat, 2014b).

Figure 2. Joint processes: the potential to use certain indicators for multiple processes.



Source: Eurostat (2014b)

Second, telecom providers also derive benefits from sharing their data, such as additional information about their market. MNOs would not be able to determine their market share amongst specific groups of travellers without having access to the total figures (Eurostat, 2014e). Thus, if MNOs collectively share their data for international travel statistics, each will gain additional information about their market, which can help them better tailor their services to different segments of travellers.

3.3.4. Legal framework

European Union regulations on personal data protection are implemented differently in each member state (Eurostat, 2014a). The main European texts include Directive 1995/46/EC or the Data Protection Directive (DPD), which regulates the processing of personal data, and Directive 2002/58/EC or the Electronic Privacy Directive (EPD), which regulates the processing of such data in the electronic communications sector. The DPD and EPD are respectively transposed into Estonian law through the Personal Data Protection Act (PDPA) and Electronic Communications Act (ECA).

A legal analysis undertaken in the context of Eurostat's feasibility study has concluded, based on the Article 29 Data Protection Working Party Opinion, that non-aggregated location data is considered to relate to an identifiable person and therefore should be treated in accordance with the DPD (Eurostat, 2014a; Borenius, 2014)[18]. The PDPA specifies that personal data may be processed if the data subject's consent has been obtained or if there exists a statutory basis for processing the data without the subject's consent (Eurostat, 2014a). According to Art 10 (2) of the PDPA, "An administrative authority shall process personal data only in the course of performance of public duties in order to perform obligations prescribed by law, an international agreement or directly applicable legislation of the Council of the European Union or the European Commission". The extent to which a given data source is considered necessary for the performance of an obligation prescribed by law remains unclear (Eurostat, 2014a). Based on the opinion of the Estonian Data Protection Authorities, the general obligation of Statistics Estonia to produce official statistics under the Official Statistics Act should be a sufficient legal basis on which to process mobile positioning data if it provides adequate justification that it cannot produce the relevant statistics otherwise (Eurostat, 2014).

In sum, while European and Estonian legislation remain ambiguous on certain aspects, there is a case to be made that mobile positioning data may be used to produce tourism statistics. It is worth noting that, if the data is to be processed by a third party (as is the case in Estonia), "they should also be subject to performing a statutory task commissioned to them under the law or a contract under public law concluded with the state statistics authorities" (Eurostat, 2014a)[19].

3.4. The Prediction of Consumer Confidence Using Social Media Data by Statistics Netherlands – a Hybrid Model

Researchers from Statistics Netherlands have explored the relationship between social media sentiment and consumer confidence in order to derive a predictive model (Daas and van der Loo, 2013; Daas and Puts, 2014). Access to the data, made up of public Facebook and Twitter messages,

[18] Article 29 Data Protection Working Party Opinion comprises representatives of European Union member states' supervisory authorities, the European Data Protection Supervisor and a representative of the European Commission (Eurostat, 2014a). Its opinions are of advisory nature (Eurostat, 2014a).
[19] For a more detailed analysis, see Eurostat (2014a, pp. 11-52) and Borenius (2014).

was purchased from Coosto, a provider of social media data and metadata. Coosto performed its own aggregate sentiment estimates (based on word combinations and emoticons) which were exported to Statistics Netherlands for further analysis. Researchers from Statistics Netherlands appeared to be familiar with the classification method followed by Coosto, but did not have access to their algorithms (Daas and Puts, 2014)[20]. Sentiment analysis was therefore conducted under a grey-box model. As in Estonia, the data was obtained through an intermediary. Thus, in terms of the taxonomy established in 2.3, while this model involves the transfer of data sets it shares features with the third party approach. The relative ease of access to the actual data (as compared to CDRs, for instance) – albeit through an intermediary, relates to the nature of data: it is made up of public messages and therefore does not contain any confidential information[21].

The results were positive, as several combinations of data were found to correlate and co-integrate with consumer confidence. The Granger causality test found that consumer confidence affected social media sentiment. The same model performed well with data from the United Kingdom. However, the report demonstrated two important methodological caveats associated with social media data. First, it is biased towards positive sentiments. Second, it is sensitive to disruptions. For instance, during Christmas and the London Olympic Games there was a prevalence of positive sentiments unrelated to any underlying increase in consumer confidence (Daas and Puts, 2014).

[20] The authors were apparently provided with the generic methodology as well as with some features particular to Coosto's tool, such as the classification of messages containing "smileys" since early 2013 (Daas and Puts, 2014).

[21] Statistics Netherlands has also explored other options for obtaining social media data. This was done through the REST API, which provides information about Twitter user identifiers. Researchers were able to gather a large number of identifiers (bandwidth constraints were overcome by using multiple accounts) and subsequently collect the messages from Dutch users (Daas et al., 2012).

4. LESSONS LEARNED

4.1. The Private Sector Can Derive Non-financial Benefits from Sharing its Data

Existing cases have shown that the private sector can derive important non-financial benefits from working with NSOs. Sharing data should be presented as an opportunity to acquire capacity and knowledge (e.g. MNOs' market share among certain categories of tourists): combined data sets can generate information which no agent could have provided on its own. Similarly, the processing tools implemented for a public-private partnership can be re-used for other purposes: past cases have shown that the tools and methodologies used to process private sector data into official statistics can and should be designed to take advantage of joint processes and produce a range of different indicators (e.g. Positium). In addition to affording opportunities for synergies, using their data for social good can provide private data holders with valuable experience in a certain area. Telefónica's work on AlertImpact has provided its research teams with experience in the field of mobility statistics which has arguably paid off through the launch of Smart Steps.

4.2. The Need for a Structured and Inclusive Approach

As stressed by Landfeld (2014) and demonstrated by the D4D challenge, PPPs should be structured by clear agreements and regulatory mechanisms, especially when data is exchanged. These mechanisms are meant to both prevent misuse of the data and promote trust between individual, corporate and public stakeholders. The D4D challenge was regulated at two levels: a) by the Commission de Protection des Données Personnelles which issued recommendations regarding anonymisation, and b) within the framework of the challenge, through the internal workgroup and DEEP (Orange, 2015b). By collaborating with privacy protection authorities, Orange removed important regulatory barriers to sharing its data. Moreover, the terms and conditions acted as a double safeguard against misuse through their eligibility requirements and restrictions of use (see Berens and Verhulst, 2015).

It is essential to engage all stakeholders and takes their concerns into account. In the short-term, mobile network operators will be willing to take maximum precautions in order to protect their proprietary information as well as the privacy of their subscribers. They will tend to either control the statistical processes (e.g. in-house processing) or transfer datasets which have undergone various manipulations (e.g. coarsening, introduction of noise) in order to maintain confidentiality. Positium illustrates how it is possible to align the concerns of individuals, telecommunication companies and statisticians, namely, by keeping all actors involved throughout the process.

The example of Telefónica demonstrates that MNOs are willing, on their own initiative, to engage in areas that matter to NSOs. This suggests that there might be numerous opportunities for statistics offices to co-operate with telecom providers on the methodological aspects of projects such as CenCell or AlertImpact in order to insure that they conform to official statistical standards.

Furthermore, it is crucial that data protection authorities are involved in the process. The Estonian case has shown that they can provide guidance when legislation is ambiguous. Private data controllers

are likely to be more confident about sharing their data once they obtain the approval of data protection authorities.

4.3. Understand the Trade-off between Granularity and Level of Access to the Data

The case of Smart Steps has illustrated that, when confident about the level of security, private data providers can link several of their datasets. For instance, MNOs can combine CDRs with the demographic information that they possess about their subscribers. If the risk of information leakage is low, which almost certainly implies a lower level of access to the data, MNOs can provide extremely valuable "enriched" data. The privacy and competitive risks associated with transferring phone logs coupled with socio-economic data are even higher than those from exporting only CDRs. Given that the latter already act as important deterrents to the physical sharing of datasets, it is likely that in the short- to medium-term, the only way to extract value form integrated datasets will be from within the MNO's systems.

Since both NSOs and MNOs are committed to the protection of personal data, in cases where a transfer of data is necessary – e.g. in order to match CDR data with household surveys or censuses-, it is unlikely to involve fine-grained data (CenCell used aggregated socio-economic data). On the other hand, coarse data carries evident limitations (e.g. might fail to capture smaller poverty enclaves).

4.4. Follow a Blended Approach

While unofficial sources of data can allow for a reduction in traditional data collection (e.g. reduce the frequency of surveys), they carry multiple caveats and cannot act as a pure substitute for existing official statistics. This may result from several factors:

- The suitability of the type of data for the chosen purpose: the example of Google Flu Trends is particularly telling in this respect. Lazer et al. (2014) argue that the Google search engine's algorithms have created a bias in the search term data through their suggestion system: search suggestions are both endogenous and exogenous to actual search queries, i.e. a particular search suggestion both depends on the amount of times it was entered but also influences how likely people are to enter the term into the search engine.
- The NSO does not possess a sufficient level of access to the data or methodology to be able to fully validate the statistics.
- The quality of the data or methodology does not meet the standard of official statistics.
- Official statistics are needed to calibrate the "new" statistics.

Hence, it is important to adopt a blended approach to the use of unofficial data sources for official statistics: traditional data sources are important for calibration and interpretability of new statistics; when possible, both should be combined. Moreover, existing datasets facilitate quality-control and validation (See Eurostat, 2014c on the use of reference data for ensuring the coherence of mobile positioning-based statistics). This is the approach adopted by Eesti Pank, which uses existing data sources (accommodation, administrative, press, etc.) to verify mobile positioning estimates (Kroon, 2012). In sum, non-official data should not be seen as an alternative to traditional sources but as a complement.

4.5. Creating a Cost-effective Model

4.5.1. The benefits of intermediaries

Intermediaries, such as Positium and Coosto, can lead to more efficient processes, as only one entity needs to invest in a processing infrastructure. This model is particularly suited for data holders which

are concerned with cost, or which simply do not wish to process the data in-house. This is the model followed by Twitter which, through its open APIs has enabled third parties to add value to its data (Bright Planet, 2013). Coosto, the intermediary, handles pooled data from different social media platforms such as Facebook, Twitter, LinkedIn and many others. This type of centralised processing structure is more efficient than a distributed and heterogeneous infrastructure. It is worth noting, however, that the use of an intermediary is not the only solution to enhance the coherence of a processing system. Data producers themselves could very well co-operate among themselves. Lorraine Stone, from Telefónica Dynamic Insights, expresses this possibility with regards to Smart Steps (Stone, 2015).

4.5.2. Pricing matters

Ideally, the providers of data or processing and analysis services should take NSOs' financial means into account when setting prices. For instance, Twitter has offered researchers access to its data through its "Twitter Data Grants", which suggests that the social networking company might be willing to provide national statistics offices with reduced prices, a necessity in the global south, where experimental studies have already demonstrated the potential of social media to address some important policy issues such as food price crises (Twitter, 2014; see also UN Global Pulse, 2014). In accordance with the recently drafted Principles for Access to Big Data Sources, both public and business interests should be taken into account when data is purchased from the private sector[22]. The data needed for the production of official statistics should ideally be priced at its marginal costs and NSOs must try to rely on as many providers as possible. As demonstrated by the Smart Steps case, private firms can earn profits through their private clients and charge NSOs a fair price without jeopardising their business.

[22] A draft of the Principles for Access to Big Data Sources was prepared for the Global Conference on Big Data for Official Statistics which took place in Abu Dhabi between 20-22 October 2015 and is available from: http://unstats.un.org/unsd/trade/events/2015/abudhabi/Draft%20Principles%20for%20Access%20to%20Big%20Data%20Sources.pdf

5. THE WAY FORWARD

The case studies have pointed to a number of lessons, which can be applied within different timeframes. NSOs can build on existing structures of co-operation in the short-term, but can also create an environment conducive to higher levels of collaboration in the long-run.

5.1. Short-term – Control Sensitive data

5.1.1. Co-operate through trusted third parties or through an in-house approach

As data-sharing PPPs are a new development, the private sector has demonstrated a large degree of caution when solicited to share their data. When data sets are physically transferred, they often lack the granularity necessary for a range of applications. The Estonian case has demonstrated how trusted third parties can reconcile the needs for detailed data and confidentiality. In-house approaches could also allow NSOs to leverage valuable sources such as combined data sets, as explained in 4.3.2. Finally, a hybrid model involving a trusted third party processing raw data from within the provider's servers could provide additional security guarantees to the data holders. For instance, the Positium Data Mediator is capable of operating from within an MNO's servers.

5.1.2. Build upon existing structures of co-operation

Despite the numerous obstacles to co-operation in the field of data, recent years have seen the emergence of co-operative structures which can be linked to the nature of specific data sources, but also to past experiences. As noted above, social media metadata is not sensitive and therefore accessible through APIs and private platforms. Yet, there are also structures for sharing CDRs, which have emerged from successful collaboration in the past. These structures often link different actors within the private sector (before working with Eesti Pank, Positium LBS had been accessing CDRs to provide services to the private sector). These can take time to build. Hence, NSOs must make the most of structures already in place. This can be done through the "third-party" approach or by exploring less sensitive sources of data.

5.1.3. Blended approaches for compiling official statistics

Presently, the degree of statistical generalisability of many non-traditional sources is not well understood (Reimsbach-Kounatze, 2015). 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 mixed approach implies that, while NSOs will continue to rely on traditional statistical methods, it will be with the aim of approving and adjusting non-traditionally compiled statistics.

5.2. Long-term – Trust, Exchange, Sustainability

5.2.1. *Develop effective sharing protocols and invest in relationships and with data holders*

Highly sensitive personally identifiable medical data is customarily shared with researchers but through strict protocols which limit the risks of individuals being re-identified (Eagle, 2009). In this respect, the procedures developed for the D4D Challenge are very promising, as they demonstrate a systematic approach to accessibility, pseudonymisation techniques and overall ethical issues. The success of PPPs in the long-run depends on the adoption of the systematic and transparent approaches

to data-sharing already in place in other sectors. These measures are essential in order to satisfy Principle 6 of the *Fundamental Principles of Official Statistics* (General Assembly Resolution 68/261, 2014), and inspire trust in the reliability and integrity of NSSs when dealing with non-volunteered data[23,24]. It is worth noting that NSOs are already subject to high professional standards and generally have a sound record with regards to confidentiality (see Struijs et al., 2014).

In addition to protecting privacy, sharing protocols should insure that the purposes for which the data is exchanged are transparent, clear and enforced. The controversial sale of anonymous traffic data by TomTom data to the Dutch police which was used to target excessive speeding illustrates the need for explicit terms and conditions determining what uses of data are appropriate under a specific agreement (see Palmer, 2011). The D4D challenge has undertaken a promising approach in this regard, by assessing each project as a whole, including its implications.

Obtaining the agreement of private partners can be a lengthy process. It is therefore crucial to invest in a relationship with data holders and build on past collaboration. Often, if is easier to access data for research purposes. Research or pilot projects can initiate a relationship, promote trust, and introduce future cooperation (Struijs, 2015).

5.2.2. Skills required by NSOs

As the private sector becomes more willing to share their data for statistical purposes, NSOs will have the opportunity to develop their own techniques for compiling statistics from non-official data, but also to use white-box testing to validate the estimations of their private partners (Reimsbach-Kounatze, 2015). In both cases, NSOs will require a new range of skills in the field of mathematics and computer science. This is likely to lead to a new understanding of statistical capacity, which will require the ability to process unofficial data sources. Unfortunately, the demand for data scientists is expected to surpass supply (Reimsbach-Kounatze, 2015), which means that PPPs might need to continue to outsource the processing and analysis of non-official data sources, due to a lack of statistical capacity, especially in developing countries. This role could also be taken on at a regional level by organisations such as AFRISTAT or Eurostat.

5.2.3. Regional statistical offices as potential intermediaries

Regional statistical offices may possess the most appropriate infrastructure to cope with the technical challenges of transferring large datasets (provided that this is the model followed). They could also act as trusted third parties and help save in co-ordination costs a) between private data holders (e.g. MNOs) and b) between NSOs. This would avoid multiplying contracts between NSOs and MNOs. This, however, would entail a harmonisation of national legal frameworks, which is likely to only occur in the longer-term.

5.2.4. From a "zero risk" approach to risk assessment and mitigation frameworks

Many data sources relevant to PPPs constitute personal data. Despite the existence of pseudonymisation techniques, it is difficult to rule out the possibility of re-identifying data subjects while preserving the data's resolution (see Eurostat, 2014e; de Montjoye et al., 2015; Narayanan and

[23] "Principle 6. Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes." (General Assembly Resolution 68/261, 2014)
[24] The OECD (2015) makes a distinction between volunteered (active subject) and observed (passive subject) data.

Shmatikov, 2008). This, however, should not justify withholding the data. Instead, steps should be taken to assess and mitigate these risks, which should be weighted against the social benefits arising from the responsible use of the data (Yakowitz, 2011). The system should be subject to audits and legislation should be updated both to introduce clarity about the appropriate uses of unofficial data sources and to "engage technological solutions to protect privacy" (Skaliotis, 2015; Eurostat, 2014a; World Economic Forum, 2014a). Certain data protection authorities are already taking steps in this direction, such as the French data protection regulator (CNIL) which, in 2010, has issued guidance on anonymisation methods (Eurostat, 2014a).

6. CONCLUSION

PPPs in statistics are a promising but underexploited solution to data gaps. PPPs can strongly contribute to mobilising the data revolution for sustainable development, as they are one of the only ways of translating private-sector data into official statistics. Although cases of successful PPPs are sparse, the examples studied have shown that the private sector is willing to co-operate with third parties, provided that their confidentiality requirements are respected. Moreover, the case studies have pointed to an important trade-off between data quality and level of access, which can be optimised in the short term through the use of third parties. Greater mutual trust resulting from a systematic approach to confidentiality and ethics can resolve this trade-off.

The study and advertising of successful as well as less successful cases, is crucial to creating confidence within both the private and public sectors. Databases such as PARIS21's innovation inventory are useful tools for centralising past experiences for the purpose of future study.

BIBLIOGRAPHY

ACAPS (Assessment Capacities Project) (2013), Call Detail Records: the Use of Mobile Phone Data to Track and Predict Population Displacement in Disasters, *ACAPS Reviews*, 12 June 2013. Available from:
www.acaps.org/img/documents/c-mobile-phone-data-for-displacement.pdf

Ahas, R., et al. (2014), *Mobile telephones and mobile positioning data as source for statistics: Estonian experiences*, Department of Geography, University of Tartu, Tartu, Estonia. Available from:
www.researchgate.net/publication/260318044_Mobile_telephones_and_mobile_positioning_data_as_source_for_statistics_Estonian_experiences

Ahas, R., et al. (2007), Evaluating passive mobile positioning data for tourism surveys: An Estonian case study, *Tourism Management* 29 (2008), pp. 469–486

Badiee, S. et al. (2004), Developing Countries' Statistical Challenges in the Global Economy. In: *9th National Convention on Statistics (NCS)*. Mandaluyong City, Philippines, 4-6 October, 2004. Available from:
www.nscb.gov.ph/ncs/9thncs/papers/plenary_DevelopingCountries.pdf

Ballivian, A. and Hoffman, B. (2015), *Public-Private Partnerships for Data (draft)*, Issues Paper for Data Revolution Consultation

Bengtsson, L. et al. (2015), Using Mobile Phone Data to Predict the Spatial Spread of Cholera, *Scientific Reports 5*, 8923 (2015). Available from:
www.nature.com/articles/srep08923

Berens, J., and S. Verhulst (2015), *Data Prizes and Challenges as Data Collaboratives – Terms and Conditions* [Online] GOVLAB. Available from:
http://thegovlab.org/data-prizes-and-challenges-as-data-collaboratives-terms-and-conditions/
[Accessed 15/12/2015]

Berlingerio et al. (2013), AllAboard: a System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data. In: *European Conference on Machine Learning* and *Principles and Practice of Knowledge Discovery in Databases.* Prague, Czech Republic, 23-27 September 2013. Available from:
www.ecmlpkdd2013.org/wp-content/uploads/2013/07/651.pdf

Bier, W. and P. Nymand-Anderson (2014), The use of non-official sources in official international economic and financial statistics. In: *The European Conference on Quality in Official Statistics. Special Session: Serving Policy Makers with International Statistics – Use of Non-Official Sources in International Statistics.* Vienna, Austria, 5 June 2014. Available from:
www.q2014.at/fileadmin/user_upload/CCSA_Q_abstract_WB-PNA_2014_clean.pdf

Blondel et al. (2013), *Data for Development: the D4D Challenge on Mobile Phone Data*. Available from:
http://arxiv.org/pdf/1210.0137.pdf

Blumenstock, J., G. Cadamuro, and R. On (2015), Predicting Poverty and Wealth from Mobile Phone Metadata, *Science*, 350 (6264), pp. 1073-1076.

Borenius (2014), *Memorandum on Legal Regulation on the use of Mobile Positioning Data in the European Union and Estonia, Finland, Germany and France*, 7 January 2014. Available from: http://mobfs.positium.ee/data/uploads/task-2/legal-memorandum.pdf

Bright Planet (2013), *Twitter Firehose vs. Twitter API: what's the Difference and Why Should You Care?* [Online]. Bright Planet. Available from: www.brightplanet.com/2013/06/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care/ [Accessed 15/12/2015]

Daas et al. (2012), "Twitter as a Potential Data Source for Statistics", *Discussion paper*, No 201221, Statistics Netherlands, The Hague/Heerlen, The Netherlands. Available from: www.cbs.nl/NR/rdonlyres/04B7DD23-5443-4F98-B466-1C67AAA19527/0/201221x10pub.pdf

Daas, P. and M. van der Loo (2013), Big Data (and Official Statistics). In: Eurostat, OECD, UNESCAP, UNECE, *Meeting on the Management of Statistical Information Systems*. Bangkok, Thailand, 23-25 April 2013. Available from: www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_4_Daas.pdf

Daas, P. and M. Puts (2014), Social Media Sentiment and Consumer Confidence, *European Central Bank Statistics Paper Series*, No 5, September 2014. Available from: www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.en.pdf

de Cordes (2015), "Low cost industrialization for PPP Analytics ?". In: OECD/PARIS21, *Access to New Data Sources for Statistics: Business Models for Private-Public Partnerships*. Paris, 18-19 December 2015. Presentation. Available from: http://www.oecd.org/std/1-Nicolas%20de%20Cordes.pdf [Accessed 03/02/2016]

de Montjoye, Y. et al. (2014), *D4D Senegal: The Second Mobile Phone Data for Development Challenge*. Available from: http://arxiv.org/pdf/1407.4885v2.pdf

de Montjoye, Y. et al. (2015), Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata, *Science*, 347 (6221), pp. 536-539.

Eagle, N. (2009), Engineering a Common Good: Fair use of Aggregated, Anonymized Behavioural Data, *IEEE Engaging Data*, Boston, MA.

Economist (2014), "Waiting on Hold" [Online] 25 October, 2014. Available from: www.economist.com/news/science-and-technology/21627557-mobile-phone-records-would-help-combat-ebola-epidemic-getting-look [Accessed 15/12/2015]

European Commission (2014), *Fact Sheet Data cPPP*. Retrieved 15 July 2015. Available from: https://ec.europa.eu/research/industrial_technologies/pdf/factsheet-cppp_en.pdf

Eurostat (2014a), Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics: Report 2. Feasibility of Access, *Eurostat contract no. 30501.2012.001-2012.452*. Available from: http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Task-2-report.pdf/3cd12ef6-939c-447b-853d-cc1552a768dd

Eurostat (2014b), Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics: Report 3a. Feasibility of Use: Methodological Issues, *Eurostat Contract 30501.2012.001-2012.452*. Available from: http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Task-3a-report.pdf/38716d6f-c2db-4ee3-92f3-0241bca55f94

Eurostat (2014c), Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics: Report 3b. Feasibility of Use: Coherence, *Eurostat Contract 30501.2012.001-2012.452*. Available from:
http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Task-3b-report.pdf/36a1a827-3c90-4e3e-bfcb-7f0534bfb06a

Eurostat (2014d), Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics: Report 4. Opportunities and Benefits, *Eurostat Contract 30501.2012.001-2012.452*. Available from:
http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Task-4-report.pdf/09dca964-034d-41da-8620-60a26f29582a

Eurostat (2014e), Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics: Consolidated Report, *Eurostat Contract No 30501.2012.001- 2012.452*. Available from:
http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf/530307ec-0684-4052-87dd-0c02b0b63b73

Eurostat (2013), Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics: Task 1. Stock-taking, *Eurostat contract no. 30501.2012.001-2012.452.* Available from:
http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Task-1-report.pdf/1e81fd37-e9d9-4c2f-ab39-3528e6c1237c

FEMA (Federal Emergency Management Agency) (2008), Emergency Support Function #6 – Mass Care, Emergency Assistance, Housing, and Human Services Annex, *FEMA*. Available from:
www.fema.gov/pdf/emergency/nrf/nrf-esf-06.pdf

Flowminder, Haiti Cholera Outbreak 2010 [Online] *Flowminder.Org*. Available from:
www.flowminder.org/case-studies/haiti-cholera-outbreak-2010 [Accessed 15/12/2015]

Frias-Martinez, V. et al. (2010), "Towards Large Scale Technology Impact Analyses: Automatic Residential Localization from Mobile Phone-Call Data". In: *4th ACM/IEEE International Conference on Information and Communication Technologies and Development*. London: IEEE. Available from:
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.207.5113&rep=rep1&type=pdf

Frias-Martinez, V. and E. Frias-Martinez (2012), *Enhancing Public Policy Decision Making Using Large-Scale Cell Phone Data* [Online], UN Global Pulse. Available from:
www.unglobalpulse.org/publicpolicyandcellphonedata [Accessed 15/12/2015]

General Assembly Resolution 68/261 (2014), *Fundamental Principles of Official Statistics*, A/RES/68/261. Available from:
http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf

Huynh, I. (2014), How is the World Bank Harnessing Big Data For Development? Presentation notes, in e-xperience 2014, Cartagena 4-5 December 2014. Available from:
http://centrodeinnovacion.gobiernoenlinea.gov.co/sites/default/files/3_exp2014_pres_bigdata_isabellehuynh.pdf

IEAG (UN Independent Expert Advisory Group) (2015), *A World that Counts: Mobilizing the Data Revolution for Sustainable Development*. Independent Expert Advisory Group Secretariat. Available from:
www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf

Institute of Business Ethics and Orange (2015), *Data for Development Senegal: Report of the External Review Panel*, April 2015. Available from:

www.d4d.orange.com/fr/content/download/43823/426571/version/2/file/D4D_Challenge_DEEP_Report_IBE.pdf

Kroon, J. (2012), Mobile positioning as a possible data source for international travel statistics. *Seminar on new frontiers for Statistical Data Collection*, Geneva, Switzerland, 31 October- 2 November 2012. Geneva, UNECE. Available from: www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP6.pdf

Kuusik, A., R. Ahas and M. Tiru (2010), The ability of tourism events to generate destination loyalty towards the country: an Estonian case study. In: Mäeltsamees, S.; Reiljan, J. (Toim.). *XVIII rahvusvaheline majanduspoliitika teaduskonverents Majanduspoliitika Eesti riikides* – aasta 2010 (156 – 175). Berliner Wissenchafts-Verlag, Mattimar.

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. Available from: http://unstats.un.org/unsd/trade/events/2014/beijing/Steve%20Landefeld%20-%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf

Lazer et al. (2014), The Parable of Google Flu: Traps in Big Data Analysis, *Science*, 343 (6176), pp. 1203-1205. Available from: http://gking.harvard.edu/files/gking/files/0314policyforumff.pdf?m=1394735706

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. Available from: www.ipinst.org/images/pdfs/IPI_Epub-New_Technology-final.pdf

Letouzé, E. and J. Jütting (2015), "Official Statistics and Human Development", *Data-Pop Alliance White Paper Series.* Data-Pop Alliance (Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute) and PARIS21. Available from: https://static1.squarespace.com/static/531a2b4be4b009ca7e474c05/t/55a1e0f9e4b03e8188399687/1436672249742/WPS_OfficialStatistics.pdf

Letouzé, E. and P. Vinck (2015), "The Law, Politics and Ethics of Cell Phone Data Analytics", *Data-Pop Alliance White Paper Series*. Data-Pop Alliance, World Bank Group, Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute. Available from: https://static1.squarespace.com/static/531a2b4be4b009ca7e474c05/t/55a1e16de4b0e8589c2f42fb/1436672365962/WPS_LawPoliticsEthicsCellPhoneDataAnalytics.pdf

McAfee, A. and E. Brynjolfsson (2012), Big Data: The Management Revolution, *Harvard Business Review*, 90 (10), pp. 60-68.

Narayanan, A. and Shmatikov, V. (2008), Robust De-anonymization of Sparse Datasets. In: IEEE, *Symposium on Security and Privacy*. Oakland, California, USA 18-21 May 2008. Washington DC: IEEE.

OECD, "Strengthening National Statistical Systems to Monitor Global Goals", *OECD and Post-2015 reflections*, Element 5, Paper 1. Paris, France: OECD. Available from: www.oecd.org/dac/POST-2015%20P21.pdf

OECD (2015), *Data-Driven Innovation: Big Data for Growth and Well-Being*, OECD Publishing, Paris.

Orange (2015a), Challenge 4 Development: Presentation [Online] D4D Challenge. Available from: http://www.d4d.orange.com/en/presentation [Accessed 15/12/2015]

Orange (2015b), "National development and population welfare take centre stage as Orange announces the winners of its big data 'Data for Development' Challenge Senegal" [Online]. Available from: http://www.orange.com/en/Press-and-medias/press-releases-2015/National-development-and-population-welfare-take-centre-stage-as-Orange-announces-the-winners-of-its-big-data-Data-for-Development-Challenge-Senegal [Accessed 15/12/2015]

Orange (2014), Orange data made available for the D4D Senegal Challenge [Online] D4D Challenge. Available from: www.d4d.orange.com/en/presentation/data [Accessed 15/12/2015]

Palmer, M. (2011), "TomTom sorry for selling driver data to police". *The Financial Times* [Online] April 28th. Available from: http://www.ft.com/cms/s/2/3f80e432-7199-11e0-9b7a-00144feabdc0.html#axzz3rLZpCAZb [Accessed 15/12/2015]

PARIS21 (2015), *A Road Map for a Country-Led Data Revolution*. Paris, PARIS21. Available from: http://datarevolution.paris21.org/sites/default/files/Road_map_for_a_Country_led_Data_Revolution_web.pdf

PARIS21 (2014), *Informing a Data Revolution Cross-Country Study*. Paris: PARIS21. Available from: http://datarevolution.paris21.org/sites/default/files/Country%20study%20write-up_sep%2022.pdf

PARIS21 (2012), *Big Data, Big Time: Statistical Capacity Development 2.0* [Online] PARIS21. Available from: www.paris21.org/node/1377 [Accessed 15/13/2015]

Pastor-Escuredo D. et al (2014), Flooding through the lens of mobile phone activity. In: IEEE, *Global Humanitarian Technology Conference (GHTC)*. San Jose, CA, USA, 10-13 October 2014. Washington DC: IEEE.

Peng et al. (2005), Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (8), pp. 1226-1238.

Pentland, A. (2012), *Reinventing Society in the Wake of Big Data: A Conversation with Alex (Sandy) Pentland*. Interviewed by Edge, 30 august, 2012. Available from: https://edge.org/conversation/alex_sandy_pentland-reinventing-society-in-the-wake-of-big-data

PPPIRC (Public-Private Partnership in Infrastructure Resource Centre) (2015), What are Public-Private Partnerships? [Online] World Bank. Available from: http://ppp.worldbank.org/public-private-partnership/overview/what-are-public-private-partnerships [Accessed 15/12/2015]

Reimsbach-Kounatze, C. (2015), "The Proliferation of "Big Data" and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis", *OECD Digital Economy Papers*, No. 245, OECD Publishing. Available from: www.oecd-ilibrary.org/docserver/download/5js7t9wqzvg8.pdf?expires=1445936759&id=id&accname=guest&checksum=E881B32E5C87C7D9296305E2B4976D5B

Rosa-Péres, E. and V. Velasco Gimeno (2012), *Automated data collection in accommodation statistics: a European overview*. In: OECD and Eurostat, 11th Global Forum on Tourism Statistics. Reykjavík, Iceland, 13-16 November 2012. Available from: www.congress.is/11thtourismstatisticsforum/papers/Session3.pdf (p.48 of the PDF)

Sabol, P. and R. Puentes (2014), *Private Capital, Public Good: Drivers of Successful Infrastructure Public-Private Partnerships*, Brookings, Washington D.C. Available from: www.brookings.edu/~/media/research/files/reports/2014/12/17-ppp/bmpp_privatecapitalpublicgood.pdf

Sanga, D. (2011), "The Challenges of Monitoring and Reporting on the Millennium Development Goals in Africa by 2015 and Beyond", *Journal Statistique Africain*, numéro 12, mai 2011, pp. 104-118. Available from: www.afdb.org/fileadmin/uploads/afdb/Documents/Publications/The%20Challenges%20of%20Monitoring%20and%20Reporting%20on%20the%20Millennium%20Development%20Goals%20in%20Africa%20by%202015%20and%20Beyond%20St12.pdf

Skaliotis, M. (2015), "Addressing Legal and Technical Challenges of Big Data in the European Statistical System: The Big Data Action Plan and Roadmap". In: OECD/PARIS21, *Access to New Data Sources for Statistics: Business Models for Private-Public Partnerships*. Paris, 18-19 December 2015. Presentation. Available from: http://www.oecd.org/std/2-Skaliotis.pdf [Accessed 08/02/2015]

Soto et al. (2011), Prediction of Socioeconomic Levels using Cell Phone Records. In: User Modelling, Adaptation and Personalization (UMAP), 19th International Conference. Girona, Spain, 11-15 July 2011. Available from: www.vanessafriasmartinez.org/uploads/umap2011.pdf

State Street Global Markets (2012), *State Street PriceStats*. Boston: State Street Corporation. Available from: http://www.statestreet.com/content/dam/statestreet/documents/pricestats/pricestats-qa.pdf [Accessed 15/12/2015]

Stone, L. (2015), A Behind the Scenes Look at Telefónica's Evolving Big Data External Monetisation Model. Interviewed by McGee-Abe. Available from: http://dynamicinsights.telefonica.com/wp-content/uploads/2015/04/Big-Data-Monetization-in-Telecoms-Smart-Steps.pdf

Struijs, P. (2015), "Access to New Data Sources: Experiences in the Netherlands". In: OECD/PARIS21, *Access to New Data Sources for Statistics: Business Models for Private-Public Partnerships*. Paris, 18-19 December 2015. Presentation. Available from: http://www.oecd.org/std/2-Peter%20Struijs.pdf [Accessed 03/02/2015]

Struijs et al. (2014), "Official Statistics and Big Data", *Big Data & Society*, April-June 2014, pp. 1-6.

Tatalovic, M. (2013), Data for Development: Revolution Kicks off in Côte d'Ivoire [Online] *SciDev.Net*. Available from: www.scidev.net/global/data/scidev-net-at-large/data-for-development-revolution-kicks-off-in-c-te-d-ivoire.html [Accessed 15/12/2015]

TechCrunch (2015), Twitter Cuts off DataSift to Step Up its Own Big Data Business. [Online] TechCrunch. Available from: http://techcrunch.com/2015/04/11/twitter-cuts-off-datasift-to-step-up-its-own-b2b-big-data-analytics-business/#.v8x67z:ufD7 [Accessed 15/12/2015]

Telefónica Dynamic Insights (2012), Telefónica Dynamic Insights launches 'Smart Steps' in the UK [Online] Telefónica Dynamic Insights. Available from: http://dynamicinsights.telefonica.com/blog/658/telefonica-dynamic-insights-launches-smart-steps-in-the-uk-2 [Accessed 15/12/2015]

Telegraph (2015), Twitter Cuts off Data Stream to British tech 'unicorn' DataSift. [Online] *The Telegraph*. Available from: www.telegraph.co.uk/finance/newsbysector/mediatechnologyandtelecoms/digital-media/11533275/Twitter-cuts-off-data-stream-to-British-tech-unicorn-DataSift.html [Accessed 15/12/2015]

Twitter (2014), Introducing Twitter Data Grants [Online] Twitter. Available from: https://blog.twitter.com/2014/introducing-twitter-data-grants [Accessed 15/12/2015]

UNDESA (United Nations Department of Economic and Social Affairs) (2015), *Global Sustainable Development Report*. New York: United Nations Department of Economic and Social Affairs. Available from: https://sustainabledevelopment.un.org/content/documents/1758GSDR%202015%20Advance%20Unedited%20Version.pdf

UN Global Pulse (2014), *Mining Indonesian Tweets to Understand Food Price Crises*, Methods Paper, February 2014. Available from: http://www.unglobalpulse.org/sites/default/files/Global-Pulse-Mining-Indonesian-Tweets-Food-Price-Crises%20copy.pdf

Upadhyaya, S. (2014), Use of Non-Official Sources for Transforming National Data into an International Statistical Product – UNIDO's experience. In: *The European Conference on Quality in Official Statistics. Special Session: Serving Policy Makers with International Statistics – Use of Non-Official Sources in International Statistics*. Vienna, Austria, 5 June 2014. Available from: http://unstats.un.org/unsd/accsub/2014docs-CDQIO/Paper-UNIDO.pdf

US OMB (United States Office of Management and Budget) (2014), *Memorandum for the Heads of Executive Departments and Agencies*, M-14-06, 14 February, 2014. Available from: www.whitehouse.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf

World Bank, *Public-Private Partnership in Infrastructure Resource Centre*, Retrieved 10 July 2015 Available from: http://ppp.worldbank.org/public-private-partnership/overview/what-are-public-private-partnerships

World Economic Forum (2005), *Data-Driven Development: Pathways for Progress*. Geneva: World Economic Forum. Available from: www3.weforum.org/docs/WEFUSA_DataDrivenDevelopment_Report2015.pdf

Yakowitz, J. (2011), "Tragedy of the Data Commons", *Harvard Journal of Law & Technology*, 25 (1), pp. 1-67.


OECD
BETTER POLICIES FOR BETTER LIVES