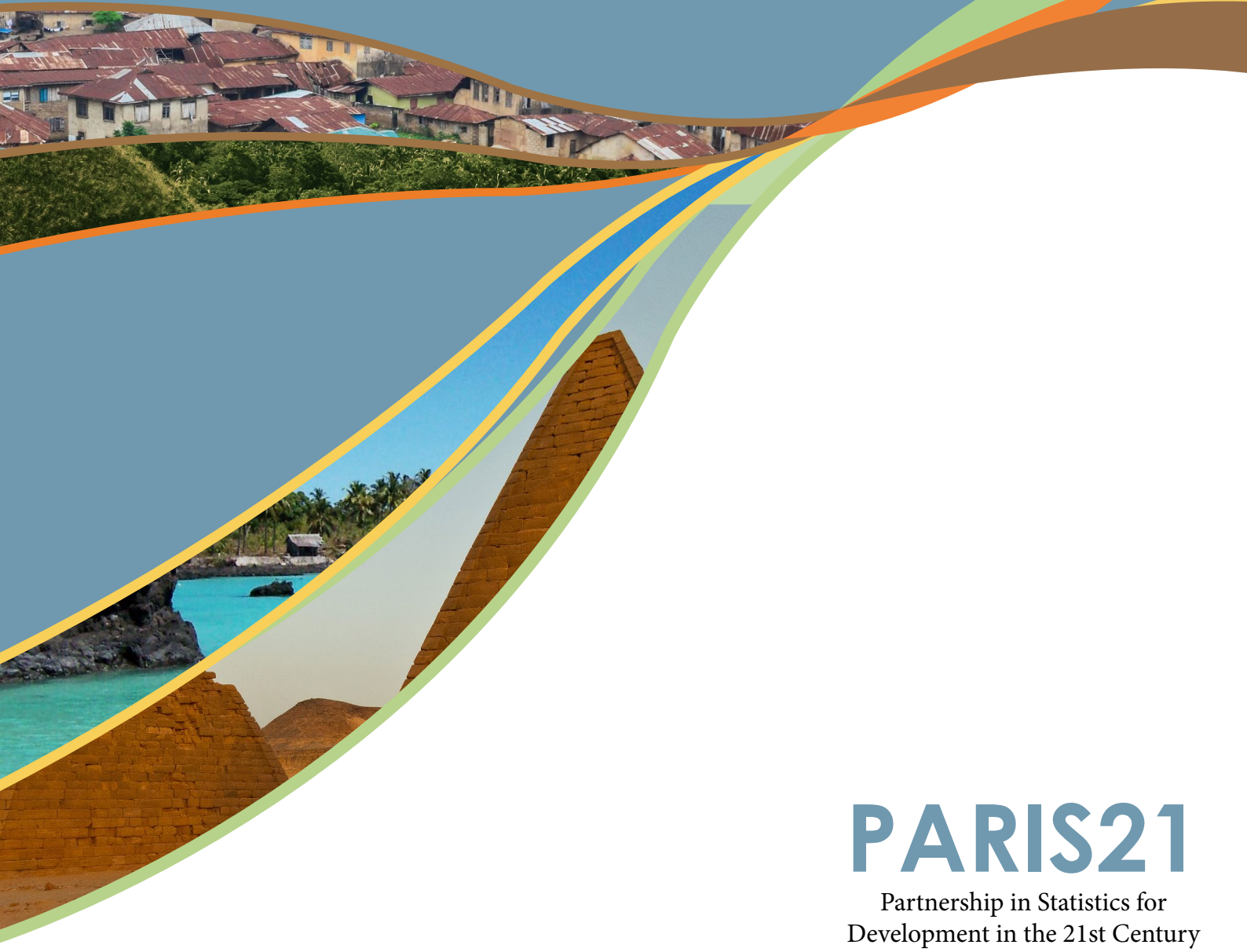
Access to New Data Sources for Statistics: Business Models and Incentives for the Corporate Sector

Thilo Klein and Stefaan Verhulst



PARIS21
Partnership in Statistics for Development in the 21st Century

Discussion Paper No. 10
March 2017

Disclaimer

The opinions expressed in this paper are those of the authors and should not be attributed to the PARIS21 partnership or its members. The opinions expressed and arguments employed are those of the authors. Discussion Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which PARIS21 works.

Acknowledgements

The authors would like to thank Gautier Krings (Proximus) and Frédéric Pivetta (Dalberg Data Insights) for their inputs as well as Marc Debusschere (Statistics Belgium), Freddy De Meersman (Proximus), Eric Anvar, Arnaud Atoch, Trevor Fletcher and Paul Schreyer (OECD) and Johannes Jütting (PARIS21) for their discussions and comments. In addition, comments from members of the Eurostat Task Force on Big Data, in particular David Salgado (INE) and Loredana di Consiglio, Fernando Reis, Hannes Reuter and Albrecht Wirthmann (Eurostat) were gratefully received. The authors are further thankful to BBVA, Google, Proximus and Telefónica for sharing their opinions.

Access to New Data Sources for Statistics: Business Models and Incentives for the Corporate Sector

Thilo Klein and Stefaan Verhulst

Abstract

New data sources, commonly referred to as "Big Data", have attracted growing interest from National Statistical Institutes. They have the potential to complement official and more conventional statistics used, for instance, to determine progress towards the Sustainable Development Goals (SDGs) and other targets. However, it is often assumed that this type of data is readily available, which is not necessarily the case. This paper examines legal requirements and business incentives to obtain agreement on private data access, and more generally ways to facilitate the use of Big Data for statistical purposes. Using practical cases, the paper analyses the suitability of five generic data access models for different data sources and data uses in an emerging new data ecosystem. Concrete recommendations for policy action are presented in the conclusions.

Keywords

Big data, official statistics, data ecosystem, business model, public-private partnership (PPP)

Résumé

Les nouvelles sources de données, désignées couramment comme le « Big Data », ont attiré un intérêt croissant des Instituts de Statistiques Nationaux. Elles pourraient compléter les statistiques officielles et plus conventionnelles, utilisées par exemple pour déterminer le progrès vers les Objectifs de Développement Durable, et d'autres objectifs. Cependant, on considère souvent que ces types de données sont facilement accessibles, ce qui n'est pas nécessairement le cas. Cet article examine les conditions nécessaires et les incitations commerciales pour obtenir des accords sur l'accès aux données privées, et plus généralement les manières de faciliter l'utilisation du Big Data en statistiques. En utilisant des cas pratiques, l'article analyse la pertinence de cinq modèles génériques d'accès aux données pour des sources et des utilisations différentes, dans un nouvel écosystème de données émergeant. Des recommandations concrètes pour l'action politique sont présentées en conclusion.

Mots-clés

Big data, statistiques officielles, écosystème de données, modèles de gestion, partenariat public-privé (PPP)

Table of Contents

Acronyms and Abbreviations

BBVA	Banco Bilbao Vizcaya Argentaria
CDR	Call Detail Records
OECD	Organisation for Economic Co-operation and Development
NSO	National Statistical Office
PDS	Personal Data Stores
PPP	Public-Private Partnership
SDG	Sustainable Development Goals
T3P	Trusted Third Party

Executive summary: Key messages

New data sources from the private sector have enormous potential to complement and enhance official statistics by, for example, providing more timely data and developing new areas for analysis. The increasing number of research projects using this data underlines its potential. However, progress to operationalize the use of such data sources in the world of official statistics is proving to be slow due, in part, to barriers in gaining access to data from private corporations. Such barriers take the form of concerns on the part of private companies about losing their competitive advantage; legal constraints concerning privacy and confidentiality of client information; and the costs of setting up the necessary infrastructure and training staff for a non-core business related activity.

Nonetheless, several business models that enable data exchange between private corporations and official statistics are emerging, including:

- in-house production of statistics
- transfer of data sets to end users
- remote access to data
- trusted 3rd parties (T3Ps)
- moving algorithms rather than data.

Each of these business models, along with their associated risks and technical and governance requirements, are examined in this paper.

Incentives for private companies to share their data include the mutual benefits accrued from working with National Statistical Offices (NSOs), the potential to develop new analytical skills, improve their reputations, generate revenue, meet regulatory compliance and demonstrate corporate responsibility. There is growing recognition among many companies – albeit slowly – of these many incentives for making data available for public good.

After examining various data sharing models and incentives, the paper makes a number of recommendations for policy actions. These recommendations suggest the need for NSOs to enter into partnerships with private providers:

1) Corporations should consider electing “data stewards” to act as focal points for data access.
2) A network to share experiences and know-how around data sharing and evidence of impact should be created.
3) A repository of case-studies that highlight innovation in sharing practices should be developed.
4) A decision tree to help assess benefits and risks of sharing corporate data should be defined.
5) A standardized safe environment for sharing data without risk of compromising customer privacy should be developed.

1. Introduction

There has been increased interest in recent years in the potential of using new data sources (often referred to as "Big Data") to complement official statistics.[1] Demand for statistics is increasing, in particular in the context of the Agenda 2030 and the global requirement to measure progress towards the Sustainable Development Goals (SDGs). At the same time, the budgets of National Statistical Offices (NSOs) are tight and pressure has been mounting to find innovative solutions to use these new kinds of data for effective and innovative public policy making.

Within this context, policy makers and others have begun looking toward new data sources (including telecommunications data, social media logs, data from sensors, satellite images, web scraping, financial transactions etc.) as sources for potential solutions to public problems. In recent decades there has been great progress in the development of technologies used to store and analyse such high-volume data sets. A lot of research (e.g. UNDESA, 2015) has also examined how this data can be mapped to existing needs, and how it can be used to provide more cost efficient and timely information.

One shortcoming of such research is that much of it begins from the presumption that this type of data is in fact readily available. However, in reality, gaining access to the data is often a formidable challenge due to privacy, confidentiality, and security concerns, as well as cross-jurisdictional regulatory incompatibilities in how data may be owned and transferred (OECD, 2013a). The need for more clearly defined business models and incentives is also required because so much big data is owned by private sector entities. As the United Nations' Global Working Group on Big Data for Official Statistics noted in its 2015 report, “the biggest challenge for most Big Data projects is the limited or restricted access to potential data sets.” The report also noted that most big data sets are owned by the private sector, often multinational companies, and that any effort to use big data for public purposes – specifically, in this case, statistics – must “build close collaboration with the private sector.”[2]

This paper builds on the outcomes of a workshop jointly organised in December 2015 by the Organisation for Economic Co-operation and Development (OECD) and the Partnership in Statistics for Development in the 21st Century (PARIS21) on “Access to New Data Sources for Statistics: Business Models for Private-Public Partnerships.” This event brought together data producers from the commercial, non-commercial and public sectors to present examples of new data sources being used for statistics, to discuss data confidentiality and privacy issues, and to identify possible business models from the perspective of the private sector in providing access to its data. To gather evidence for the paper, the authors conducted a survey (see Annex 1) among 70 data-producing companies from a wide range of sectors, including social media, telecommunications, banking, online retailers and software vendors. The survey aimed to better understand the companies’ motivations and perceived risks when engaging in partnerships for official statistics by asking whether they provided data to official statistical bodies and, if not, what prevented access from being granted. Disappointedly, he response rate to the survey was very low with only four completed surveys

[1] It is important to acknowledge, however, that not all new data sources are big nor are all big data sources new.
[2] http://unstats.un.org/unsd/statcom/doc15/2015-4-BigData.pdf

returned from the companies contacted. Nevertheless, some interesting insights can be drawn from the survey results.

This paper aims to examine potential ways of addressing some of the barriers to increased sharing. Building on existing literature (see Box 1), it presents possible business-sharing models that could contribute towards a set of guidelines, principles or protocols to enable greater use of private data for the public good. Section 2 presents the different data sets and actors that could provide new sources to statistical agencies. This "data ecosystem" consists of various data producers and users, as well as governance mechanisms that determine how data is collected and shared; understanding this ecosystem is essential to understanding both the opportunities and challenges inherent to the use of private data by public entities. Section 3 considers some of the specific risks and incentives involved when sharing data, and Section 4 presents various sharing models that could help both public and private entities navigate the previously discussed ecosystem. In conclusion, Section 5 outlines policy recommendations for private companies and public policy makers seeking to use the vast stores of privately held data for statistical modelling and analysis.

Box 1: Existing literature on public and private sector cooperation in data sharing

The paper draws from, and contributes to, an emerging strand 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 (see Reimsbach-Kounatze, 2015, for an early overview). Existing studies include a recent paper by Robin, Klein and Jütting (2016), who categorise four generic types of public-private partnerships (PPPs) for statistics with a focus on the use of non-official data sources for national statistics and public policy. Ballivian and Hoffman (2015) drew lessons for PPPs for data from past and existing multi-stakeholder agreements and offer a taxonomy of risks and benefits of data sharing. Two reports specific to telecom data for official statistics include de Meersman et al. (2016), who report on a "win-win" partnership between a telecom operator and Statistics Belgium, and Eurostat's (2014) "*Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics*". The Eurostat report draws on several 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. Finally, a much broader overview on the issue is given in the World Economic Forum (2015) report on "*Data-Driven Development*" and Landfeld's (2014) discussion paper on the "*Uses of Big Data for Official Statistics*". The "*Data-Driven Development*" report offers a comprehensive account of the challenges faced in harnessing privately held data for development as well as a number of solutions. Landfeld (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.

2. New data sources for statistical agencies

Questions concerning who owns and has access to data sources are intrinsically complex and often create inherent tensions between public and private entities. Private (i.e., corporate) entities often use data to maximize their individual profits and market share, while public entities (e.g.,

governments and civil society organisations) seek to use data to help increase the overall welfare of a given population.

2.1 Data ownership

This paper is primarily concerned with data controlled by a private company or other entity (private data) (OECD, 2013b). A company is typically said to "own" data when its clients have explicitly consented to the use of data collected during the provision of a service, or sometimes because an earlier owner of the data has sold it or transferred ownership rights. However, in many cases (e.g. when data is generated via scanners used for retail purchases or through Internet cookies), clients may not have explicitly consented to ownership of data by the private entity, but ownership is nonetheless in effect established. In other cases, data may have been generated by individuals but cannot be said to belong to them, e.g. in the context of telecoms network data where it belongs to the telecom operator rather than the customer.

Data ownership is therefore often a contentious and confused issue, even when ownership is clear from a legal perspective. For instance, many private data sources rely on the use of public or shared infrastructure auctioned by public bodies, such as telecoms infrastructure or wave bandwidth. Determining when such data should remain private and when it should be shared more widely can be difficult and is particularly relevant for governments or other public statistical organisations that seek to leverage private data for the public good. Questions arise about what data to use, how to gain access to it, and whether access should be gained through the market (e.g. payment) or through legislation and policy that impose certain public sharing or use requirements on private stakeholders.

2.2 Data sources

Data can come from a number of sources, and privately used data has already been used around the world to generate public insights. Annex 2 highlights some key data sources, as well as the types of data available from each of these sources. Some of the most widely used data sources are telecom operators, satellite companies and social media platforms. These platforms today occupy a central role in our everyday lives and offer the potential for a wide a range of insights into social relations, economic activity, population movements and various other aspects of public life.

When considering data ownership and models for data transfer or sharing, it is perhaps easier to stake a public claim to data generated on public infrastructure or infrastructure originally built with public funds (e.g. telecoms networks or GPS). A growing amount of data, however, is generated by private-sector entities like retailers and financial services companies. Such data, including point-of-sale retail data and credit card information, is no less potentially useful as a source of statistical insight, and should not be overlooked by governments or statistical bodies seeking to supplement publicly held information.

It is important to note that privately held data may contain certain regional or other biases. This may be particularly true of data in developing countries, where network penetration rates may be low and where other sources of data may reflect an urban, regional, educational or income bias. To take just one example, the Nakumatt Supermarket in Kenya, which was the first retail chain to offer a

loyalty system and collect data on consumer purchases, may at first glance seem like an attractive source of information on consumer habits; however, the company has only about a 35% market share, most of which comes from urban or semi-urban areas. The remainder of the market is covered by a long tail of small shops (Muganda et al. 2014). Although less pronounced, a narrow sample and regional bias contained in such data can also be seen in more developed countries, and may limit the effectiveness of privately held data as a tool for general statistical information or modelling.

3. Incentives and risks in sharing corporate data

Companies that see no clear reason to share their data (or that are overly focused on the risks of sharing) are less likely to do so and, as a result, the public is poorer or at least missing out on an opportunity. Greater clarity on the incentives and risks of sharing is thus essential. The various risks and incentives posed by data sharing are examined below (Verhulst et al. 2017).

3.1 Incentives

Corporations agree to provide access to their data for different reasons, depending on the context in which the data is being requested or shared, the question access to their data may answer, and the corporate and legal culture of the firm. Different corporations also have different views regarding the expected benefits and risks from sharing their data. In other words, when firms do extend themselves and share their data they seek to satisfy a variety of motivations.

Below are six types of incentive or motivation observed from existing case-studies where corporate data was made accessible.

3.1.1 Reciprocity

Corporations may share their data with others for mutual benefit, especially where gaining access to other data sources may be important to their own business decisions. Others may reciprocate to give back what was taken from individuals and society-at-large.

Example 1: Data pools under the Accelerating Medicines Partnership (AMP)

Data pools created under the Accelerating Medicines Partnership (AMP) allow genetic and molecular data to flow between 10 private pharmaceutical companies, the United States National Institute of Health, and the Food and Drug Administration. The initiative aims to overcome fragmentation in the pharmaceutical industry and improve innovation in drug therapy, ultimately allowing pharmaceutical companies to find new drug targets and reduce wasteful repetition of testing found when companies work in silo. Such improvements in private-sector efficiency highlight the reciprocal benefits gained via data collaborations, where private companies are motivated to share data with both private and public partners in their best business interests.

3.1.2 Research, recruitment and insights

A corporation opening up its data may generate new answers to particular questions providing otherwise non-extractable insights. Just as with open source, sharing data (and, in some cases, algorithms) enables corporations to tap into data analytical skills (often free labour) available outside their own company. External users may examine the data in new ways, and use the skills and methodologies not readily available internally. Sharing may also create the opportunity to identify and hire (or in some cases retain) valuable talent. In addition, these insights may enable companies to identify new niches for activity and to develop new business models.

Example 2: The Banco Bilbao Vizcaya Argentaria (BBVA) Innova Challenge

The Spanish bank BBVA Innova Challenge allows participating researchers and developers to access BBVA's Big Data. With this data, they create apps and compete for awards based on the usability and innovation of their digital products both within and outside the company. The winners of the Innova challenge have not only spurred innovation in the public sector, creating an app that predicts overcrowding in city buildings for example, but also have identified ways BBVA can more effectively support their customers. By opening up their corporate data through this initiative, BBVA have supported research and innovation for the public good and their private commercial interests.

Example 3: Population estimates through mobile phone data in Belgium

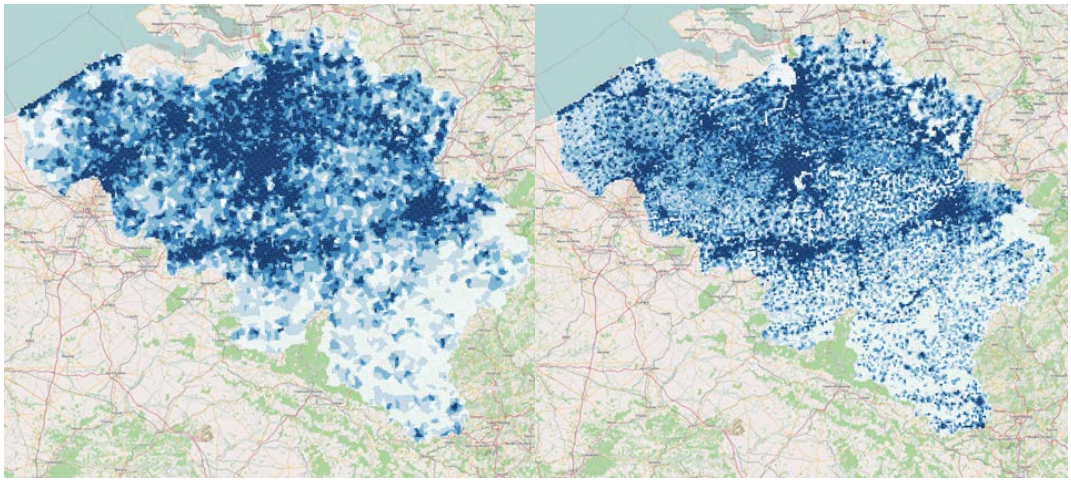
In December 2015, Statistics Belgium, Eurostat and Belgium's major network operator, Proximus, launched a project to jointly assess the commercial and statistical uses of mobile phone data. The case demonstrates a business model based on a "win-win" partnership in which statistical institutes and mobile network operators combine their data and resources to obtain results that neither could have achieved on its own.

Proximus, like many other operators, is increasingly aware of the value of the data 'exhaust' of the functioning of the mobile network. In addition to the traditional use of call detail records (e.g. billing operations) as well as network optimisation signalling data, a growing number of operators now see the commercial potential of the data they hold. In order to turn this data into valid and accurate information, operators can benefit substantially from both the statistical skills and access to representative, geo-coded micro-level data sets that are available in NSOs.

The project started by estimating the population, resident population, place of work, and advanced to the measurement of commuting, tourism, labour mobility and migration. The analysis combined mobile phone data with statistical data sets which makes it possible to validate one against the other and create additional information. In order to do so, the geographical units of mobile phone data (the area covered by each of the 11,000 mobile network cells) have to be converted in the standard 1 km² area of statistical data sets, and vice versa (much higher precision is foreseen in future studies).

The initials results presented in Figure 1 below are published in de Meersman et al. (2016). The graph shows population densities derived from mobile phone counts at 4am on Thursday 8 October 2015 (left) and the 2011 population census (right). The Pearson correlation between these two data sets is 0.85, a clear indication that mobile phone data are able to provide a valid and accurate measure of population density.

Figure 1: Population density per km² based on mobile phone data (left) and 2011 census (right).



The partnership business model pioneered by Statistics Belgium and Proximus provides incentives for both statistical institutes and mobile network operators to pool resources and data in an equitable way, producing valuable results for both partners, while legal, operational and business issues seem manageable. Furthermore, collaboration can evolve into a long-term relationship essential for the production of statistical time series.

3.1.3 Reputation and public relations

Sharing data for public good may enhance a firm's corporate image and reputation, potentially attracting new users and customers. It may also offer an opportunity to gain (free) media attention and increase visibility among certain decision makers and other audiences.

Example 4: Orbital Insights contribution to the World Bank's Measuring Poverty from Space project

Orbital Insights is a data analytics company that provides products to both private businesses and non-profit entities that enables data-driven decision making. It recently collaborated with the World Bank in its Measuring Poverty from Space project, providing satellite data and geo-analytics to help the World Bank track poverty around the world. By focusing on such data collaborations, Orbital Insights has generated significant investor interest, recently raising USD20 million from venture capital groups and In-Q-Tel, an American non-profit firm specializing in investments for intelligence products. By addressing needs from both the private and public sector, Orbital Insights

has gained a reputation for high-level analytics and widened their market options.

3.1.4 Increasing revenue

Corporate data is not necessarily always opened for free. Under some conditions, corporate data may be offered for sale, generating extra revenue for firms.

Example 5: Telefonica Smart Steps' sharing of anonymous aggregated data

As part of Telefonica's Insights arm, Telefonica Smart Steps releases anonymous aggregated mobile data from its network of operators and analyses trends and patterns useful for its clients. Though some of its clients are from the business sector, others involved in public infrastructure use Telefonica's data analysis to improve policy and service delivery. By sharing its data, Telefonica has expanded its role as a telecommunications provider and capitalized on the value of its data to aid public problem-solving.

3.1.5 Regulatory compliance

Sharing data can help corporations comply with sectoral regulations and become more transparent and trusted. In addition, many corporations generate data to meet regulatory compliance. Sharing and using data in a responsible manner (see below regarding some of the risks of opening up data) to benefit both the public and private sectors increases the value of the investment made to collect the data for a narrower purpose.

Example 6: Opening up data from the Employer Information Report Surveys

All companies based in the United States are required to file Employer Information Report (EEO-1) Surveys through the Equal Employment Opportunity Commission which collects employment data on race/ethnicity, gender and job category. Some companies choose to release this information to the public. Open Diversity Data collects and showcases all companies that release their EEO-1 information and those that do not. Apple, Cisco, Dell and Google (among others) all release the data contained in their EEO-1 filings on their websites, ensuring that the diversity in their workforce is transparent, promoted and maintained. By sharing the public data used for regulatory compliance, companies allow people to more broadly scrutinize discrimination within their institutions, thereby encouraging greater diversity in their workplaces.

3.1.6 Responsibility and corporate philanthropy

Sharing corporate data achieves many of the goals sought by traditional corporate social responsibility or philanthropy. Under this incentive, a company derives value from socially responsible behaviour not just because of the positive image such activity produces, but because opening up data can also improve the ecosystem within which the business operates (Stempeck, 2014). A classic example of such an ecosystem-supporting responsibility is a company that contributes data to help improve education, which could eventually improve the labour pool from which they hire staff.

Example 7: Data sharing for social good, Nielsen and Feeding America

Since 2010, the sales and marketing analysis group, Nielsen, has released food pricing information to Feeding America to assist with their advocacy and food monitoring efforts. Together, they created the Map the Meal programme which analyses food insecurity in America. Nielsen entered this partnership as part of its Nielsen Cares initiative whereby its data is used for social good, contributing to their corporate social responsibility efforts. In such a way, Nielsen is able to use their existing data and expertise to benefit the wider society and improve their corporate image.

3.2 Risks and potential harms of sharing corporate data

Whether corporations are willing to share data depends not only on the (perceived) benefits and motivations but also on the (perceived) risks and harms of doing so. Sharing corporate data poses a number of risks that may harm the entities involved in the exchange, as well as those intended to benefit from the exchange. Risk mitigation strategies tackle these risks and potential harms by, for example, limiting data access to specific, pre-approved uses. They may also "bring the algorithm to the data", an arrangement in which private-sector data sets never leave corporate databases but are processed and analysed using external algorithms that may, for instance, be hosted in the cloud**.**

Data risks exist throughout the data value cycle--from collection, to analysis and processing, to use. Risks are often the result of technological weaknesses (e.g. security flaws); individual and institutional norms and standards of quality (e.g. weak scientific rigor in analysis); legal confusion or gaps (e.g. weak or no privacy provisions); or misaligned business and other incentives (e.g. companies seeking to push the boundaries of what is socially appropriate). While there are common elements across these risks, it is useful to examine them by separately considering each stage of the data value cycle.

When risks are not addressed at the initial stages of the value cycle (e.g. when data is not "cleaned" at the collection stage) they may accumulate and lead to additional risks downstream (e.g. making flawed inferences from the data analysis due to inaccurate data). This is why it is important to take into account potential risks not just at the points of access and sharing, but also at the data collection stage. Some of the most common risks are presented below.

3.2.1 Data collection

At the data collection stage, risks include:

a) collecting inaccurate, old or "dirty" data affecting data quality and determining data output
b) collecting unauthorized data or intrusive collection from individuals and organisations – including the use of flawed consent mechanisms
c) incomplete or non-representative sampling of the universe – ignoring "data invisibles" or population segments with a limited data footprint.

Box 2: Focus on data quality

The usefulness of data is directly correlated to its quality. Data quality is a technical term that encompasses a number of properties, each of which should be evaluated (where possible) by any agency or organisation seeking to use privately held data for public purposes. Failure to evaluate the quality of data may result in erroneous analyses and conclusions that could lead to ineffective — or worse, harmful — public policies.

Some of the key elements of data quality include:

- Integrity: at a most basic level, quality data is free (or at least relatively free) of errors. Errors in data can themselves take at least two forms. They can result from incorrect information (e.g. information that may have been sourced incorrectly from the field) or information that may have been entered erroneously (data entry errors). A lack of integrity in data can also result from data corruption, which can itself be attributed to a number of sources, including viruses, malware, or damage to physical media or other technologies used to store or transmit data.

- Freshness: quality data is data that is fresh or up-to-date. Indeed, this is one of the primary potential benefits of private held big data sets, many of which are regularly updated on a real-time or near-real-time basis (e.g. from the ongoing activities of users whose data is being collected). Such regularly updated data can help policy makers make more immediately relevant decisions based on actual on-the-ground conditions.

- Richness: by combining and merging information from different data sets, policy makers and others can create richer data. Rich data is more descriptive and informative, and thus more useful. It allows for the collation of different and apparently disparate sources of information to reveal new insights about a population, the economy and other public priorities. Rich data leverages the reality of data proliferation to add a further and essential element to the notion of data quality.

3.2.2 Data analysis and processing stage

At the data analysis and processing stage risks include:

a) insufficient, outdated or inflexible security provisions creating data vulnerabilities or breaches
b) aggregation and correlation of incompatible data sets
c) poor problem definition or research design, flawed data modelling or use of biased algorithms.

Example 8: Google Flu Trends

Google Flu Trends provides a good example of a data collaborative that was weakened by flawed data modelling. The 2013 initiative attempted to provide real-time predictions of flu prevalence by analysing Google's search terms. However, in its inaccurate algorithm which was too broad and mistakenly identified seasonal search terms, like "high school basketball", as flu predictions. As a result, Google Flu Trends' predictions were off by 140 percent at the peak of the 2013 flu season.

3.2.3 Data use stage

At the data use stage, risks emerge from misinterpretation of the data, the possible re-identification of ostensibly anonymized data, and flawed decisional inference. Within the context of shared corporate data, additional risks at the use stage include:

a) lack of interoperable cultural and institutional norms and expectations
b) lack of data stewardship at both ends to ensure the responsible use of personally identifiable information as it travels across use cases and sectors
c) improper or unauthorized access to shared data
d) conflicting legal jurisdictions and different security levels.

Example 9: inBloom dismantling due to public concerns

Public concerns surrounding the collection and use of personal data, particularly of children, led to the dismantling of the non-profit organisation, inBloom. Initiated in 2011 through USD100 million in seed money from the Bill and Melinda Gates Foundation and the Carnegie Corporation of New York, inBloom aimed to store, clean and aggregate student data for states and districts, making data available and standardized for approved third party applications and software designed for educators. Backlash came from increased privacy concerns from parents and educators over the collection, use and storage of personally identifiable information on students, causing many partners to back out of initial agreements. inBloom ended its services in April 2014.

3.2.2 Harms

When risks are not mitigated – either technically, organisationally or through policy – business that share corporate data can face several potential harms. These include:

a) criminal or civil legal investigations and/or regulatory fines
b) loss of regulatory licenses, standards and certifications
c) reputational and industrial damages impacting competitive positioning and advantage including:
    - drop in share price and/or increase in cost of capital
    - customer attrition rates
    - lower employee recruitment, productivity and retention
d) an overall increase in operating expenses.

Ultimately, any use of private data sets for public purposes that is erroneous or otherwise faulty carries a risk of undermining public trust not only in technology and big data, but more generally in public institutions and governance. Such risks must be taken particularly seriously in an era of already diminishing public confidence in governance institutions and various forms of expertise. Presenting a public policy as indisputably “accurate” because it is supported by data overlooks the inherent fallibility of data and algorithmic methods and, in so doing, does a disservice to the real benefits such methods can bring to crafting better and more effective public policy.

3.3 Data governance

How risks and benefits are evaluated—and addressed--depends to a large extent upon whether there is an existing data governance context in which the data is shared. The governance context (i.e. the rules and policies that determine how data is collected, stored and shared) determines the appropriateness of sharing data and the best way to mitigate risks. The data governance ecosystem should be seen, broadly, to include a variety of state and non-state actors, as well as other forces and elements (e.g. standards or code, which can also play a powerful though non-traditional governance role). When considering ways data can be better used for statistical or other public purposes, there are three aspects of the governance context that are particularly important to understand.

3.3.1 Regulatory constraints

At the time of writing, there exists little guidance regarding how companies should use private data sources. Historically, telecom operators and satellite providers tend to consider citizen-generated data as their own property rather than as a public good. The thinking is that, as it is their investment in infrastructure that has enabled the data collection, hence the data must be theirs too. In practice, this attitude facilitates the use of data stored in private databases. However, any use of such data should come with at least two important limitations or considerations. First, as data become a private good, more thought should be given to financial flows and remuneration accruing to the private companies, and questions should be asked about whether some of those flows should be directed toward the original producers of the data (i.e. the consumers), for instance in the form of lower prices. Second, access to data must generally be limited for research and development purposes.

In the context of leveraging private data sources, the most important existing regulatory constraints relate to the protection of privacy. The privacy regulations often involve provisions on how data is hosted, as raw data should not be exported from the premises of the data producers. This does equally apply to any algorithms used as any transformations of the raw data should only be applied within the local infrastructure.

3.3.2 Technical context

Formal policies and laws are not the only elements of governance that matter in the data ecosystem. Over 15 years ago, Lawrence Lessig (1999) argued that, in modern information and networked systems, code or technology itself functions as a kind of law. As such, in considering the aspects of governance that may influence or even determine use of private data for statistical purposes, it is important to examine the variety of technical elements that may play a role.

The manner in which private data is hosted and shared is of particular importance, and determines what and how much a statistical or other public organisation may do with it. For instance, in a **decentralized aggregated data model,** data owners host and retain control of the data and provide aggregated data on demand to third party organisations (e.g. a statistical organisation). In a **decentralized raw data model**, ownership and storage of data remains the same, but third party organisations have access to a far wider variety and quantity of data. In both **models**, the owners could also provide cloud-based analytical tools and services to a third party organisation.

Finally, organisations may also choose to operate a centralized model, where aggregated or raw data is stored solely on the data owner's servers. Such data can be transferred upon request or need to a third party organisation which essentially becomes the new owner of the data, meaning it is obligated to perform its own analytical functions.

There are certain advantages and disadvantages to each of these models. The centralized and decentralized raw data model in particular raises greater privacy concerns. The decentralized aggregated data model has fewer privacy implications (since the data is aggregated and thus effectively anonymized), but it could reduce the flexibility of the third party organisation to use the data and thus reduce its usefulness. These and other concerns must be weighed when making technical decisions about how to share data. It is important to recognize that technical decisions and technical architecture play an important role in determining the usefulness of data and how it may be used.

4. Generic data access models

This paper has thus far considered various sources for private data, as well as the incentives for and risks posed by sharing such data. But assuming a company or other entity concludes that the incentives outweigh the risks and decides to share private data it may have collected for the public good—then what form should such sharing take? This section examines five different models that can be considered for data sharing and access.

- in-house production of statistics
- transfer of data sets to end users
- remote access
- trusted 3rd parties (T3P)
- moving the algorithms.

Each model is examined along four dimensions: the technical environment within which access takes place; the governance environment; the associated risks; and the types and purpose of data that might be best suited in each case. Examples are also presented for each model.

Each of these models has advantages and disadvantages. There is, in particular, a recurring trade-off between privacy and cost (or difficulty) of implementation. The most expensive and technically difficult to implement models (e.g. in-house production of statistics) also suggest a higher degree of privacy and security protections for data; at the same time, such walled-gardens of data offer less flexibility and variety, and thus possibly less potential for new and innovative insights. When deciding which model to implement, statistical or other organisations seeking to harness the potential of big or private data will need to consider these trade-offs to maximize the potential of sharing while minimizing its risks. As we have seen in Section 3, the risks can be quite significant.

4.1 In-house production of statistics

The in-house production of statistics model is, in many ways, the most conventional or standard model. It is used by the majority of statistical agencies today and, as such, comes with a known set of risks and opportunities. On the positive side, the model allows the data owner to maintain total

control over the generation and use of its raw data. User privacy can be protected through de-identification and generated indicators can be aggregated sufficiently to be considered safe for sharing. From a safety or security point of view, the in-house production of statistics is the most preferable option.

The model also has certain limitations however. In particular, there is an inherent limit on the types and nature of data available. In this model, data producers (the same as the data users) do not benefit from external expertise or from the vast amounts of data being generated every second by external agencies, companies, governments and other entities. In addition, the data owner needs to have both the in-house infrastructure and technical capability to produce these statistics on a regular basis. Given the associated costs, such a model can only be considered viable when the data owner is able to monetize the production of these statistics through a commercialised system.

Technical environment

The in-house production of data (on any scale) usually requires a certain level of internal technical capability, as well as the resources to manage and store large data sets. Companies or other entities that produce their own data are often well funded, independent agencies with well-defined needs; their technical teams are well-trained in performing a variety of support and maintenance functions. Depending on the scale of the data operation, such entities may also require in-house data scientists or experts—a resource-intensive proposition.

Governance

Governance is relatively simple in the case of in-house production of data, as the data producer and user are the same entity. Large organisations may nonetheless require agreements or guidelines for how data is shared internally between different groups and departments. In addition, data production and data sharing needs to conform to relevant government guidelines and policies (e.g. surrounding privacy and how data is collected).

Risks

With this model, the data producers control the underlying raw data and, although it requires significant investments from the producer in terms of software tools and analytical resources, it offers greater safety in terms of privacy as no third party is involved. In terms of regulation, it is necessary to know whether the data producers are using individual data for commercial purposes. For example, telecom operators have a mission to offer telecommunication services between individuals. However, adjustments to individual contracts may be required (e.g. opt-in or opt-out) to legally and contractually allow the telecom operators to use the data generated by their clients while making calls, accessing websites or sending text messages.

Type of data and purpose of use

The in-house production of data model would be most appropriate in the following contexts:

- Skills and use cases: telecom operators having strong technical skills and capable of connecting with potential end users. The latter requires an ability to understand use cases and a significant experience and investment in social responsibility and commercial goals beyond the core telecom objectives.

- Types of data: limited need to mix telecom data with additional or third party data sources. For instance, use cases requiring a map of mobile or internet penetration, or charts to assess the flows of mobile money transfers.
- Governance: limited technical capacity and governance of the overall Big Data ecosystem is required.

Example 10: Telefonica's Smart Steps project

Telefonica's Smart Steps project is based on the anonymised data produced by the mobile users of Telefonica. This data represents billions of events recorded each day and covers user consumption habits, mobility and social network, which are captured through phone calls, SMS messages and data connection logs. This data is used to produce insights such as demographic segmentations or commute patterns which are then shared with third parties in an aggregated format, extrapolated to represent the global trends of the market.

Source: http://dynamicinsights.telefonica.com/smart-steps/

4.2 Transfer of data sets to end user

In this model, data sets are moved directly from the data owner to the end user. The model gives the end user significantly more flexibility on how the data is used. In general, raw data is de-identified, sampled and sometimes aggregated to avoid possible re-identification. Efforts to de-identify need to ensure that data cannot be re-identified by crossing it with external data. Because de-identification is never absolute, even when the most sophisticated anonymizing techniques have been deployed, data in this model is generally released to a limited number of end users, under strict non-disclosure and data usage agreements that help ensure a level of control and privacy.

Releasing such granular data is best suited for research purposes, where detailed, individualised information is often required for analysis. Given the limited number of data transfer releases and the overall fit with research goals, this model tends to be used in a framework where the incentive of the data owner is to receive free research and insights on its own business, while perhaps allowing researchers to publish data results.

Technical environment

In this case, the data producer willing to share a data set provides direct copies of the relevant underlying raw data. The data user then develops the algorithms required to compute relevant aggregates and statistics to achieve its own objectives. A hypothetical example would be a Ministry of Health using data from electronic medical reports and telecom data to predict the spread of an epidemiologic disease. The data is provided by the private stakeholders in charge of each activity, allowing the mixing of different sets and sources covering various topics. The complexity comes from the skills required to handle large volumes of data from different standards and formats.

Governance

In regulated industries, such as the telecom sector, while explicit approval of consumers is typically not required, regulators need to be informed of transfers of raw data to end users. Upon review, regulators may subsequent limit any financial flows between the end user and the data producer or prevent transfer of individual data from data producers to any third party, including end users. In other sectors and regions, individual consent is often required on the part of individuals to have their data shared while others might accept implicit compliance.

Risks

There is a significant increase in operational risks for data producers that transfer data sets outside their premises. Consequently, data producers can be reluctant to send individual data to any third party, including possible end users. Safety issues include firewalls, different security and accreditation levels, backup procedures and infrastructure, secured connections, management processes, etc. Any data owner transferring individual data to a third party could potentially be held liable for any subsequent breach of privacy or misuse of that data.

Types of data and purpose of use

Transferring data sets to end users would be most appropriate in the following contexts:

- Skills and use cases: telecom operators that have invested less in technical skills and the local ecosystem. Such business model set up a transactional or commercial relation between end users and data providers, for instance telecom operators. There is limited need for the telecom operator to clearly understand and map the landscape of potential end users.
- Types of data: limited need to integrate third party data sources unless the end users play the role of data broker and integrator. In most cases, such a business model allows telecom data to mix with the end user data.
- Governance: technical capacities of the ecosystem to deal with large volumes of data and security requirements, e.g. privacy issues, secured transmission and hosting of data.

Example 11: Orange's Data For Development challenges

The best known examples of this model are the two Data For Development (D4D) challenges organised by Orange. De Montjoye et al. (2014) described the data set in detail. The goal of the contests was to develop value-added applications ranging from disease monitoring to public transport improvement for developing countries using telecom data. The winners were invited to implement their applications in practice. The contest captured the interest of hundreds of research groups across the world and, to share data in a secure manner, Orange only shared portions of data and always in a highly aggregated form. One of the data sets released represented the number of calls between cell towers; another represented the rolling two weeks mobility information for a sample of the individuals at tower level and another represented mobility information for another sample at district level.

Sources:
http://www.d4d.orange.com/en/Accueil
http://arxiv.org/pdf/1407.4885v2.pdf

4.3 Remote access

In the remote access model, data owners provide full data access to end users while maintaining strict control on what information is extracted from databases and data sets. In this model, personal identifications are anonymized, but no coarsening is actually made on the data. The data does not leave the premises of the data owner; rather, the end user is granted secured access to analyse the data and compute the relevant metrics. The end user is then permitted to extract only the final aggregated metrics once the data analysis has been completed. This method is often used in research, in specific partnerships between the data owner and a group of researchers, under very strict non-disclosure and data usage agreements. Strict monitoring of the input and output traffic on data storage devices is carried out to ensure no data is removed. The main incentive in this type of model is that users benefit from free research resources on their data.

Technical environment

The remote access model has similarities with the in-house production model. A key difference, however, is that the operation of connecting remotely to the raw data is outsourced to an external party. This requires stability in terms of connection to the infrastructure of the data producers, although it demands less in terms of specific skills from the data producer. The model allows for various third parties to access the same data sources, with only limited involvement and investment from the data producers. It might, however, still require some technical investments on the side of the data producer, such as (i) server capacity to host the raw data, algorithms and outputs; (ii) maintenance capabilities to secure sustainability and continuation; and (iii) development skills to ensure the anonymization of individual data.

Governance

The remote access model does not require any physical transfer of data but it is still likely to require regulatory approval, especially in certain sensitive industries like healthcare. Contractually, it may also be more complex if the data producer uses a third party, e.g. a technical party, to manage remote access. Despite additional intermediary costs, there are obvious benefits, such as specialization and specific cross-industry knowledge with a potentially large decentralized network of contacts and relations between end users, data producers and third parties. All of this limits the systemic risk of failures but increases the distribution of responsibilities and the risk of leaks from the weakest links.

Risks

With remote access there should be no increase in operational risks as only aggregates can be transferred. However, there are still potential risks from anyone attempting to identify groups of people or individuals by leveraging the aggregates from the third parties accessing the data. In addition, competitors of the data producers may seek to derive commercial and strategic insights from the aggregates advertised or shared by the third parties accessing the data.

Types of data and purpose of use

The remote access model would be most appropriate in the following contexts:

- Skills and use cases: best suited in the context of an early stage market where there is no clear use case and relevant end user. In this case, the third party is in charge of matching data providers and potential end users.
- Types of data: the possibility of mixing a large number of data sources, which depends on the data brokerage capacities of the third party. Depending on the third party, this model allows rich insights to be derived and possibly address complex use cases.
- Governance: a governance structure in the ecosystem exists and there is strong third party activity in the market. Such governance might be defined externally (e.g. regulator or NSO) or internally (e.g. the third party data aggregator).

Example 12: Data for Good initiative by Real Impact Analytics ; Flowminder in Nepal

The Data for Good initiative by Real Impact Analytics, a Belgian Big Data start-up, is an example of the remote access model. In this project, the company accesses telecom data within the secured environment of the operators and produce valuable insights for local authorities and non-governmental organisations concerning urban development or the spread of disease. These insights are exported via mobility maps or recommendations on where to act to control the spreading of a disease, without any personal information being released. Similarly, after the April 2015 earthquake, the organisation FlowMinder was given access to telecom data in Nepal directly at the operator's premises in order to produce maps of population displacements. These maps were shared with the United Nations to coordinate emergency response.

Sources:
https://realimpactanalytics.com/en/data-for-good
http://www.flowminder.org/case-studies/nepal-earthquake-2015

4.4 Trusted 3rd party

In the Trusted 3rd party (T3P) model, neither the data owner nor the data user support the security burden of hosting the data themselves. Instead, both parties rely on a trusted third party to host the data and enable secured access to the data source.

The data is anonymized in the sense that personal identifiers are protected by hashing techniques. In addition, the end user does not have direct access to the raw data. Instead, end users must make a request for reports or other intermediate results to the T3P, which ensures protection of the data.

This model is often facilitated by commercial contracts allowing the data owner to monetize its data. In addition, the T3P method can be well suited for regulatory initiatives, e.g. where a country requires its telecom operators to store copies of mobility data in a vault hosted by a T3P which is linked neither to an operator nor to the government. This data can then be accessed, for instance, in the case of a natural disaster.

Technical environment

The T3P model requires a technical infrastructure, such as a trusted cloud, in which each data producer can securely store its data. This requires a large data storage capacity and stable connections. Underlying data cannot leave the premises of the data producers, which implies that some anonymization and a level of aggregation needs to be carried out within the infrastructure of the data producers. Such a model facilitates the mixing of different data sources.

Governance

This business model is highly centralised around a single trusted third party who is subject to a number of laws and regulations designed to protect privacy and ensure the security of data in general, especially personally identifiable information. Standards or accreditation agencies may also be required to certify the third party's credentials.

Risks

The T3P model is high risk and requires significant investment to establish appropriate checks and balances on the T3P. In a sense, the risks of sharing with an external party are doubled when a T3P gets involved, as the number of external parties handling data has similarly doubled. The T3P model implies a certain trade-off between cost and convenience, on the one hand, and safety and security on the other.

Types of data and purpose of use

The T3P model would be most appropriate in the following contexts:

- Skills and use cases: a strong set of technical skills and connections to the potential end users at the level of the party hosting the data is needed.
- Types of data: a large range of options to mix different types of data is required, allowing rich insights to be derived and possibly complex use cases to be addressed.
- Governance: significant governance is needed as data is managed outside the premises of both the data provider and user.

Example 13: Emergence of Personal Data Stores, OpenPDS project

An emerging example of the T3P model can be found in the emergence of Personal Data Stores (PDS). These are companies that store personal data of individuals from different sources (telecom companies, social networks, mobile applications) and, with the individual's consent, provide secured access to the data to T3Ps wishing to use it. The secured access does not permit direct access to the raw data, but will rather handle user queries such as "has this user been in neighbourhood x in the last two weeks?" In this model, user identifiers are always anonymized and the PDS has a personal contract with each individual, meaning this happens at the initiative of the individuals themselves as a way to monetize their own data.

A practical example of such an application is OpenPDS, developed by the Massachusetts Institute of Technology (MIT) Media Lab. OpenPDS stores personal data in a secured environment and allows only access through a question and answer system. The questions and answers are

organised in such a way that it prevents re-identification of individuals from the answers themselves. In addition, individuals sharing their data in this way have control over which end users can access their data. This tool is currently being tested in a number of research experiments.

Source: http://openpds.media.mit.edu/

4.5 Moving algorithms rather than data

In this model, shared algorithms allow the reuse of software by several private data owners wishing to perform similar analytical functions on one or several data sets. For example, such a model may be effective in a case where several national telecoms operators wish to estimate population density (or other population patterns) based on their collective data. The data sets from different operators do not need to be necessarily merged. Instead, while the analytical functions performed on each data set may be identical, the data sets themselves can remain separate and under separate control. Results can be arrived at independently by each operator, and the aggregated results can later be merged to arrive at a comprehensive national or regional analysis.

Technical environment

The algorithm model addresses some of the challenges faced by companies in developing internal competencies for managing and analysing large data sets. Overall, this model may require less technical competence (than for example a pure in-house model), or at least the investments and HR resources to perform such analysis can be shared. Investments still need to be made in: (i) training packages to use the algorithms; (ii) standardization of the connectors/connections to the underlying raw data; and (iii) standard interpretations of the outputs and capacity to translate them into insights and actions. All these components are key to securing the existence of a standardized product which can then be moved from one data producer to the next.

Governance

Data producers are responsible for running algorithms and providing access to aggregated results. They may be required to manage relations with the parties developing and using the algorithms. The data producer may also provide support and maintenance of the software.

Risks

The algorithm business model limits risks as most of the analyses are run within the infrastructure of producers for each data set. Data producers also control access to the outputs of the algorithms, thereby further limiting the potential for leakage and other risks.

Types of data and purpose of use

The algorithm model would be most appropriate in the following contexts:

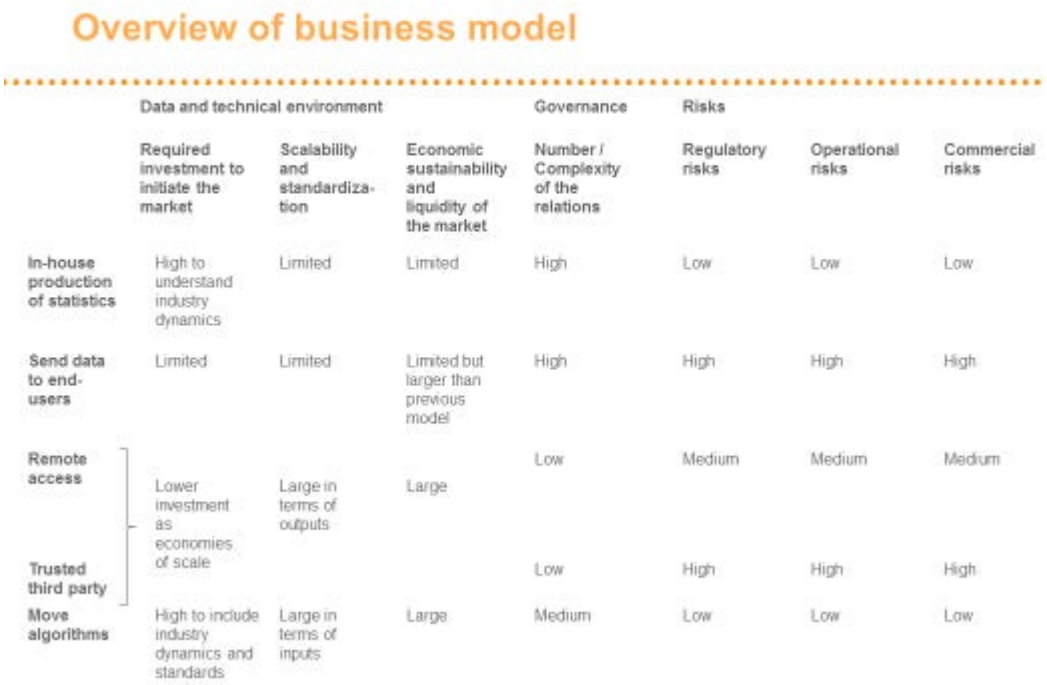
- Skills and use cases: technically strong service providers would be necessary as this business model needs to establish a connection between algorithms and data. In addition, there is a need to build visualisation or decision-making tools.

- Types of data: as algorithms need to be standardised in terms of data inputs, this imposes constraints on the complexity of the data that can be used.
- Governance: there is no need for a strong governance of the ecosystem in this case as the algorithms are decentralized and publicly available.

Example 14: Open Algorithms project

The Open Algorithms (OPAL) project provides an open platform and ready-made algorithms that allow private companies to run predefined algorithms autonomously in their own secure environments and output only the aggregated results. This method provides a unified open-source platform to the private data owner community allowing them full control over the process. Algorithms can thus reduce the workload for private companies as the software is already written and they also prevent external parties' direct access to the actual data source.

Source: http://opalproject.org/



5. Conclusions and recommendations for policy actions

This paper starts from the premise that access to new data sources, in particular data collected and stored by private organisations, can bolster NSOs in their efforts to provide reliable and actionable insights. This potential is evident in a number of ways.

Data from private sources can complement or replace the existing data and approaches being used by NSOs. When used in a responsible and methodologically sound manner, corporate data can:

- Increase the scope and breadth of NSOs' insights around existing and new metrics (including those related to the SDGs) which are otherwise hard to measure using existing data sets or costly to develop. For example, mobile and transport data can provide mobility and activity patterns to expand insights on economic well-being;
- Improve the quality and credibility of NSOs' data and analysis by merging existing NSO data sets with new ones that may validate or complete them;
- Enable more timely and more regular data analysis than data collected by NSO surveys. For example, social media streams can be used to "nowcast" social developments (in contrast to census data).
- Enable NSOs to leverage new (big data) methodologies and tap into new talent thus allowing them to become more innovative;
- Decrease NSO costs of analysis by limiting or complementing more expensive surveys.

As a result NSO users and those setting public priorities and policies can gain new and more granular insights into existing problems, in turn allowing them to become more agile, effective and innovative in problem solving and decision making. The granularity of large data sets is of particular importance and can facilitate policy makers' efforts to target specific communities or locations that are otherwise overlooked. In addition, private data may have a freshness or immediacy (particularly when it is regularly updated) that allows policy makers to set policies based on near-real-time conditions, rather than on the picture often offered by data sets from more conventional sources such as public censuses or surveys.

Despite these clear potential benefits, NSOs do not always find it easy or simple to access private data. A number of challenges exist - most prominently that private companies may be unwilling to offer access to their data due to the risk of losing competitive advantage, privacy and legal considerations, or the potentially high cost of setting up the necessary technical infrastructure and skill sets to enable data sharing.

Nevertheless, this paper argues that a number of incentives exist for private companies to work with NSOs. These include the possibility of gaining new analytical skills, reputational improvement, generating additional revenue, enhancing regulatory compliance and demonstrating corporate responsibility. The growing prevalence of data sharing initiatives suggests that companies are recognising the incentives and advantages, albeit slowly, of making their data available for the public good.

Widening access to corporate data sets will require substantial efforts by a variety of stakeholders. The move from data shielding to data sharing will require a cultural shift in the way companies and governments manage their data. To enable real change, a multi-pronged approach is needed, including:

a) Nomination of Data Stewards: A key challenge to engage with companies to access corporate data sources is the current lack of clarity as to which individuals are tasked or have the authority to consider data requests from third parties. In order to streamline data

collaboration, corporations should consider creating the role of "data stewards" to act as focal points for handling requests to access corporate data. These data stewards would be responsible for responding in a more effective and consistent manner to external demand for data, as well as coordinating with the various data actors within a company.

b) Creation of Network(s) of experts: Networks should be created to share experiences and know-how on the sharing and use of private data sources. Networks of experts and mentors could also offer lessons derived from past experiences.

c) Repository of case studies: To broaden the understanding of existing practices – both the successes and failures – and to inspire more experimentation, a repository of detailed case-studies should be created highlighting innovative sharing practices, detailing what has worked, and why.

d) Responsible Data Decision tree: To ensure that accessing corporate data sources does not harm individuals and organisations, a responsible data decision tree should be developed to help assess the benefits and risks of exchanging data.

e) Common Trusted Sharing Environment: To avoid the burden associated with establishing a safe environment in which data can be securely shared without risk of compromising customer privacy, a common trusted data sharing environment should be created. This environment could be set up by private companies themselves, or by using the services of a T3P that would serve as an industry standard leader and setter.

In conclusion, these recommendations suggest the need for NSOs to enter into partnerships with private providers. Those partnerships should vary depending on the data use and take account of the characteristics of the generic data access models described in the paper.

References

Ballivian, A. and W. Hoffman (2015), "Public-Private Partnerships for Data: Issues Paper for Data Revolution Consultation", World Bank, Draft, January 2015, http://data.worldbank.org/sites/default/files/issue-paper-financing-the-data-revolution-ppps_0.pdf

de Meersman, F. et al, (2016), "Assessing the Quality of Mobile Phone Data as a Source of Statistics", conference paper presented at the European Conference on Quality in Official Statistics, Madrid, 31 May-3 June 2016, https://ec.europa.eu/eurostat/cros/system/files/assessing_the_quality_of_mobile_phone_data_as_a_source_of_statistics_q2016.pdf.

de Montjoye, Y. A. et al (2014), *D4D-Senegal: The Second Mobile Phone Data for Development Challenge*. https://arxiv.org/pdf/1407.4885.pdf

Eurostat (2014), *Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics: Consolidated Report,* Eurostat, Luxembourg, http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf/530307ec-0684-4052-87dd-0c02b0b63b73

Landfeld, S. (2014), "Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues", discussion paper presented at the 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, http://unstats.un.org/unsd/trade/events/2014/beijing/Steve%20Landefeld%20-%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf

Lessig, L. (1999), *Code and Other Laws of Cyberspace*, Basic Books Publishing, New York.

Muganda, D.A., R. Otuya and M. Waiganj (2014), "Effect of Customer Loyalty Schemes on Competitiveness of Supermarkets in Kenya", *European Journal of Business and Management*, Vol. 6/16, published online, pp 155-164, http://www.iiste.org/Journals/index.php/EJBM/article/view/13373/13631

OECD (2013a), *The OECD Privacy Framework*, OECD Publishing, Paris, http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf

OECD (2013b), "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value", *OECD Digital Economy Papers*, No. 220, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k486qtxldmq-en

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, Paris, http://dx.doi.org/10.1787/5js7t9wqzvg8-en

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

Stempeck, M. (2014), "Sharing Data Is a Form of Corporate Philanthropy", *Harvard Business Review,* July 24, https://hbr.org/2014/07/sharing-data-is-a-form-of-corporate-philanthropy

UNDESA (2015), *Global Sustainable Development Report,* United Nations Department of Economic and Social Affairs, New York, https://sustainabledevelopment.un.org/globalsdreport/2015

UNECE (2013), "Classification of Types of Big Data", UNECE Task Team on Big Data, http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data.

Verhulst, S. et al (2017), "Data Collaboratives", http://datacollaboratives.org

World Economic Forum (2015), *Data-Driven Development: Pathways for Progress*, World Economic Forum, Geneva, http://www3.weforum.org/docs/WEFUSA_DataDrivenDevelopment_Report2015.pdf

Annex 1. Survey on Data Access

Company: Contact name: Contact email:	

1. Data access

1. Does your company already provide data for use by official statistics/research bodies (what category of data, to whom, how often, in what format, costs etc.?)
2. If not:
    a. Do you have plans to do so in the future? (When, what category of data, to whom, in what format, costs etc.)
    b. What is preventing your company from granting data access? (Data confidentiality/privacy issues, etc.)

Which data types does the company handle or generate?

Data Type (UNECE classification)	Check and comment if applicable:
1. Social Networks (human-sourced information)	
1100. Social Networks: Facebook, Twitter, Tumblr etc.	
1200. Blogs and comments	
1300. Personal documents	
1400. Pictures: Instagram, Flickr, Picasa etc.	
1500. Videos: YouTube etc.	
1600. Internet searches	
1700. Mobile data content: text messages	
1800. User-generated maps	
1900. E-Mail	
2. Traditional Business systems (process-mediated data)	
21. Data produced by Public Agencies	
2110. Medical records	
22. Data produced by businesses	
2210. Commercial transactions	
2220. Banking/stock records	
2230. E-commerce	
2240. Credit cards	
3. Internet of Things (machine-generated data):	
31. Data from sensors	
311. Fixed sensors	
3111. Home automation	
3112. Weather/pollution sensors	
3113. Traffic sensors/webcam	
3114. Scientific sensors	
3115. Security/surveillance videos/images	
312. Mobile sensors (tracking)	

3121. Mobile phone location	
3122. Cars	
3123. Satellite images	
32. Data from computer systems	
3210. Logs	
3220. Web logs	

2. Company perspective

1. What demands do they have from statistical agencies regarding data access
2. What are their view of statistical agencies and data demands
3. What needs to change/improve in statistical agencies regarding data access requests

3. Business model scenario

If they do/plan provide access to their data, which of the business models do they use?

1. Transfer of raw data
    a. Outsourcing of national statistical office functions. Activities which are typically conducted by an NSO are outsourced to a contractor on grounds of efficiency.
    b. Transfer of private data sets to the end user: Involves the physical transfer of databases to 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.
    c. Transfer of private data sets to a trusted third party for processing and/or analysis. This model involves an intermediary analysing and disseminating (and possibly processing) data.
2. No transfer of raw data
    a. In-house production of statistics by the data producer. In this type of arrangement, data is processed and analysed by the data producer, within its systems, which minimises any confidentiality risks.
    b. Use of Open Algorithms (Orange proposal)
    c. Remote access by a 3rd party to data sources (Real Impact Analytics model)

4. Motivations for providing data access

1. What are the actual or possible motivation and business arrangements?
2. What is in it for the company?
3. What level of financial compensation?
4. What are the perceived potential reputational gains?
5. Does the possibility exist to gain experience from NSOs in exploiting data?
6. Would the threat of regulatory obligation enable data access?

5. Data access agreements

1. What kind of legal agreement (if any) is in place with users of data?

2. For multi-national companies would there be a preference for setting up agreements with countries or with regional regulatory bodies

6. Data requests

1. Have they received data requests from national agencies, researchers etc. (from whom, what type of data, how often etc.)
2. Who in the company make decisions on data access?

7. Technical issues

What are the key technical issues and challenges in making data available?

Annex 2. Data Sources

Telecom data

Telecommunications companies generate a number of different data types, such as Call Detail Records (CDR), that documents the details of a telephone call or other telecommunications transaction (e.g. text message). The record contains various attributes of the call, such as time, geo-localisation, duration, completion status, source number and destination number. Telecom data also includes information on network activity, internet usage, mobile money transfers and recharging of SIM cards.

The complexity of accessing telecom data comes from:

a) privacy - the need to protect privacy of the clients generating the data
b) business - the need to protect the strategic and commercial insights from the telecom operator, as well as reputational concerns
c) legal - the need to ensure compliance with regulatory and legal constraints and requirements
d) technical - the need for telecom operators to provide at least the raw CDRs, which requires establishing a server and connection and ensure maintenance, especially in the context of a continuous data feed.

Satellite data

Satellite data can be defined as either publicly available images or privately owned images.

Publicly available images are generally of medium resolution and high quality with different levels of correction and availability; available as long-term time series; and have a larger number of spectral bands available. They are also available free-of-charge.

Privately owned images are generally of a high quality and high resolution with a limited number of visible bands, and are refreshed on a regular basis which improves their usefulness in identifying dynamics and patterns of underlying behaviours. Such images are not usually free-of-charge.

In addition to images, satellites can record data such as tracking ships, aeroplanes, temperature and other measurement data.

The complexity of accessing satellite data comes from the need:

a) to secure the knowledge to transform images into insights and identify zones and details of interest
b) to ensure high computational power and data storage capabilities necessary for processing images and storing insights
c) for auxiliary data to support knowledge extraction, e.g. climate data, terrain models, etc.

Social media logs

Social media platforms such as Facebook and Twitter permit public access to anonymised extract files of their user messages via Application Programming Interfaces (APIs). These data can be used

with text editing software for applications such as sentiment analysis or using geo-location information for population movement.

In some cases, Facebook or Google disclose information such as social contacts to assess credit scoring in the context of microloans and microfinance institutions in emerging countries or the basic analytics from Google to assess the spread of pandemic diseases based on keyword searches.

Banking transaction data

Commercial banks store large amounts of financial transaction data going back many years. These data remain the property of the banks and are not leveraged to address social or public questions due to customer confidentiality legislation.

Retailer data

Retailer data are widely used by official statistics bodies or specialised resellers in the form of scanner data. The data is purchased as a regular source of statistical information in particular for price statistics.

PARIS21

Partnership in Statistics for
Development in the 21st Century

Discussion Paper No. 10
March 2017