United Nations
Global Working Group on Big Data for Official Statistics
Task Team on Access and Partnerships

Deliverable 1&4 (merged): Good practices for data access and partnerships

DRAFT
12/10/2015

This document is based on analyses of practices in data access and partnership provided by statistical organisations and study of literature. It provides a list of goods practices (not exhaustive) that are relevant when establishing partnerships in relation to Big Data project including considerations when acquiring data access from Big Data providers. It makes part of the deliverables of the Task Team on Access and Partnerships (TTAP).

Table of Contents

Objective

Identification of good practices of data access and other types of partnership, covering the most promising Big Data sources and other types of collaboration, in particular those already identified by the GWG, for countries in all stages of economic development. Where possible, recommendations are formulated. This includes recommendations on building partnerships aimed at developing tools and skills that can facilitate data access.

Introduction

At the first meeting of the Global Working Group on Big Data it was agreed that there is a wide range of issues concerning the use of Big Data as a complementary source to official statistics. Therefore, it would be more efficient to part the work by forming eight task teams, each one addressing topics such as satellite imagery, skills and training, the Sustainable Development Goals, privacy and access and partnerships, among others. The Task Team on Access and Partnerships produced a work program covering the following topics: good practices of data access and partnerships, principles of data access, and templates for memorandums of understanding. This note is concerned with the good practices of data access and partnerships.

The partnerships can be broadly defined covering various forms of partnership from data access to data analysis, however to underline the importance of establishing cooperation with big data providers, the partnership in data access is highlighted separately. Furthermore, the nature of cooperation between data providers and other partnership partners can be quite distinct: with data providers is about data access and ensuring privacy protection, and with other types of partnership is about technologies, skills and methods used to process and analyse the data.

There is not only need to collaborate with institutions and organizations of the private sector, but also with the academy and research centres, among others, is crucial for the effective exploration of new data sources, methodologies and technologies aimed at complementing, improving and innovating the traditional way in which official statistics are produced and communicated to the public.

Partnerships can be categorised according to subject (e.g., data access, knowledge sharing) and to type of partnering institution (e.g., mobile phone provider, research institute). Where possible, recommendations are formulated and templates provided.

In this document, the Task Team proposes a set of good practices with the objective of sharing them with other national statistical systems to support the agreement, establishment and continuity of partnerships and associations for developing Big Data projects. These partnerships can be made between two or more institutions, and they may be of different nature: private firms domestic and foreign; other public agencies from central (federal) or local governments; universities, research centres national and international; international organizations; national statistical offices from other countries, among others.

Consideration

According to the Terms of Reference of the Task Team: “The objectives of the TTAP are to facilitate access to Big Data sources for official statistics and facilitate forming partnerships with other public and private organisations in order to work with Big Data… and … This should be done in a way that reflects a mutual understanding with partners of what is reasonable to expect from each other, by respecting each other’s position, role, aims, business model, social responsibilities, limitations and possibilities.”

The same document states that “Access to Big Data sources and forging partnerships with other public and private organisations in order to work with Big Data is becoming ever more important to national statistical systems (NSS) for fulfilling their mission in society. The NSS should collaborate rather than compete with the private sector, in order to advance the potential of official statistics. At the same time, the NSS should remain impartial and independent, and invest in communicating the advantages of exploiting the wealth of available digital data to the benefit of the people. Building public trust will be the key to success.”

For National Statistical System, big data sources are considered part of external (data) sources including administrative data. The fifth fundamental Principles of Official Statistics state that “Sources and methods for data collection are appropriately chosen to ensure timeliness and other aspects of quality, to be cost-efficient and to minimise the reporting burden for data providers”[1]. The good practices of this fifth principle includes (1) Facilitating the provision of data by countries, (2) Working systematically on the improvement of the timeliness of international statistics, (3) Periodic review of statistical programmes to minimise the burden on data providers, (4) Ensuring that national statistical offices and other national organisations for official statistics are duly involved and advocating that the Fundamental Principles of Official Statistics are applied when data are collected in countries

[1] http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.asp

These sources have the following characteristics: (1) the data is collected originally for specific purposes, and not necessarily designed to fit for certain statistical framework, (2) the data may contain data items/variables that may or may not be useful for statistical use, therefore, it is necessary to select and filter them, (3) the data is normally managed and owned by other public institutions or private sectors, therefore accessing it may require ad-hoc arrangement (if it is not part of legislation), (4) the data is collected regularly as part of day-to-day activities, therefore it is very timely and low cost (compared to survey), and (5) data checking and verification on input raw data may or may not exist.

Based on the result of survey conducted by UNSD/UNECE in 2014 on organizational context and individual projects of Big Data, it is cited that while IT, skills, legislation, and methodology are recognised as challenges for statistical organisations to use Big Data sources, the respondents argued that limited access to those sources is a major road block. To certain degree, the big data are collected, owned and managed by private enterprises that have defined and established specific terms of data use and reuse with their clients/customers (e.g., obligation not to share individual information with other third party). These terms may prohibit direct access by statistical organisations, and it may require collaboration or partnership on data pre-processing, processing and analysis between NSS and data providers/intermediaries themselves. One of the examples is the guidelines on the use of Call Detailed Records (CDRs) issued by GSMA on the protection of privacy in the use of mobile phone data for responding to the Ebola outbreak[2]. These guidelines had been reused to estimate mobility of the population during 2015 Nepal earthquake using mobile phone data[3].

Furthermore, the 2014 survey concluded that many organizational aspects of big data, such as quality, confidentiality, access and partnerships, should be discussed and tackled at global level due to their commonalities in many countries/regions. Considering that multinational companies are indeed the major providers of big data (e.g., Twitter has almost global coverage, mobile phone companies normally have subsidiaries in other countries), therefore, establishing global partnership would benefit multiple NSSs. This is, in fact, the broad goal of UN Global Working Group on Big Data for Official Statistics Task Team on Access and Partnerships. However, this specific document would focus on the identification of best practices of big data access and partnerships covering the most promising big data sources and various types of collaboration, including the tools and skills required for facilitation of data access.

In general, various types of partnerships (in data access, data pre-processing, data modelling, data analysis, or provision of data infrastructure, etc.) can be formed between national institutions producing official statistics and:

- Private companies
- Research and academia organizations
- Civil society organizations
- Other government institutions
- International organizations

[2] http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2014/11/GSMA-Guidelines-on-protecting-privacy-in-the-use-of-mobile-phone-data-for-responding-to-the-Ebola-outbreak-_October-2014.pdf

[3] http://www.worldpop.org.uk/nepal/

Access and Privacy

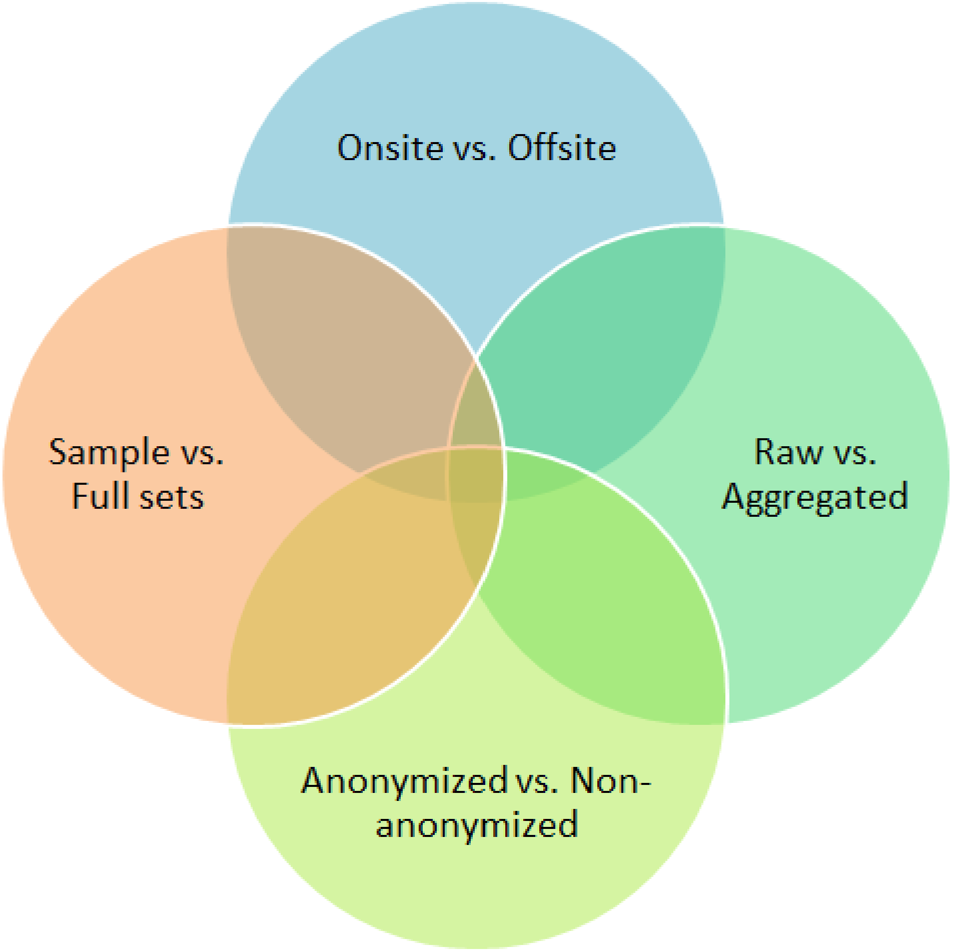
In the era of Big Data, statistical organisations are not the main data providers and must rely on other providers (including other national agencies or private companies). Even though, the sixth Fundamental Principles of Official Statistics states that “Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.”, some attempts to access Big Data source failed mainly because of concerns on privacy and individual data protection. This can be explained partly due lack of knowledge of fundamental principles of official statistics by non-statistical organisations and perception of low technical capabilities to handle the sensitive data itself. In this context, the Organization for Economic Cooperation and Development (OECD recommended that “research funding agencies and data protection authorities should collaborate to develop an international framework code of conduct covering the use for research of new forms of personal data, particularly those generated via network communication. This framework, built on best practice procedures for consent from data subjects, data sharing and re-use, anonymisation methods, etc., could be adapted as necessary for specific national circumstances”[4]

In many countries, data privacy protection is governed by various national laws, regulations, guidelines and code of practices, which are not necessarily internationally comparable. However, many of these national laws appear to follow the general recommendation of the OECD Guidelines on the “Protection of Privacy and Trans-border Flows of Personal Data[5]” (updated in 2013). This guidelines aim to protect against the disclosure of data that may pose a danger to privacy and individual liberties whether in the public or private sectors and in either electronic or paper form. The guidelines should be expanded to include the provision of access of personal data for research purposes (including for official statistics). See the example of CDR data access in the battle against Ebola outbreak in the introduction.

[4] http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf

[5] http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

Illustration of different level of data access taking into account different aspects



Discussion on access to the data can be simplified with the following aspects: 1) Location: where the data is stored (inside data provider facility or not), 2) Level of details: how many variables are made available (raw or aggregated data), 3) Identification of individuals: whether data is anonymised or not, and 4) Coverage: if the data is a sample set or full set. Based on the experience so far, it is unlikely that data provider would share the detailed dataset due to concern on data privacy and possible security breach. Therefore, by having combination of aspects above, it would facilitate data access itself. As an alternative, is to make the data provider a partner in data analysis, therefore there is no need to share the detailed dataset, and the data provider would provide the final statistical products.

In addition, statistical organisations as third party organisations receiving data should consider the risk of data breach, as there are different levels of infrastructure, experience protocols to store, manage and process full-raw-disaggregated-non-anonymized data within their own location. If the data in the hand of statistical organisations is to be lost, hacked, stolen or leaked, they can be held liable, depending on the agreement.

Particular attention should be directed to the use of web scraping to obtain data (e.g., price data) because even though the information is available publicly (through internet) and public can access the data (using web browser), downloading data excessively through web scraping may violate the website's terms of use, and/or infringe certain intellectual property rights, such as copyright or database rights pertaining to the information on the webpage. . In this context, the regulatory framework applicable in each may vary from country to country.

Tools and skills required for facilitation of data access

There are many software tools available for processing, analysing and disseminating the results of Big Data projects. In general tools should allow getting access to information from systems running on the Internet; perform powerful mathematical calculations; and, produce visualizations and analyses. For these tasks APIs, utilities and libraries, and statistical and mathematical applications will be required.

However, analysts should be aware that these tools may be different according to the different characteristics of the types of Big Data. The following examples illustrate that the proper choice of tools will be made according to nature of the data one is working on:

If data are embedded in texts, web pages, and so on, the selection would point to:

- Web scraping tools
- Text mining software

If data are non-structured:

- No-SQL databases capable of managing large amounts of electronic documents contained on file systems
- Scripting tools and languages to manipulate texts (capable of filtering, cutting, transforming, replacing, etc.)

If data are very large in volume

- Distributed file systems,
- Parallel enabled processing tools and languages

Just having the software tools and knowing how to operate them will not be enough; depending of the nature of the data sources and the intended objectives and results, knowledge of techniques and methodologies under the umbrella of the field of data science, like text mining, sentiment analysis, signal processing, probability models, machine learning, statistical learning, data mining, database, data engineering, pattern recognition and learning, visualization, predictive analytics, uncertainty modelling, data warehousing, data compression, etc. will be necessary to deal with Big Data along with other skills on statistics, mathematics and in specific the knowledge of the business.

Given the diversity of tools and skills that can be applied, the necessity will arise to create multidisciplinary working groups within National Statistical Systems for exploiting Big Data sets. Nevertheless, this opens up possibility to establish partnerships with other institutions (e.g., research institutes, data providers) so that the available tools and skills from those institutions can be utilised to access and process the raw data. The NSO can then use the pre-processed data for further data processing and integration into official statistics.

Partnerships: Good practices

A partnership can be defined in legal or non-legal terms, i.e. “a legal relation existing between two or more persons contractually associated as joint principals in a business “, or “a relationship resembling a legal partnership and usually involving close cooperation between parties having specified and joint rights and responsibilities”. In the wider context of Big Data for official statistics, partnerships include the relationship of two or more organizations conducting actual collaborations without necessarily resembling a legal partnership. Most of the following practices are the product of hard earned experiences by different partners who at some point agreed on partnerships for a wide range of purposes, including those related to the production, integration, dissemination and conservation of official statistics.

Generic steps of establishing a partnership

1. Identify the need for a partnership
2. Find a suitable partner
3. Identify common grounds of professional interest and mutual benefits
4. Write agreements
5. Develop a work program
6. State the rights and responsibilities
7. Measure and follow-up
8. Maintain communication
9. Termination of a partnership

Consideration when establishing partnership

1. Identify the need for a partnership to advance projects that require the conjunction of different parties that are willing to share skills, expertise, training, a hands in experience and/or advice, where everyone involved benefits. Projects looking to break new grounds, of innovative nature are good candidates for creating partnerships.
2. Find a suitable partner looking for sustainable partnerships. Look for a partner or partners sharing similar interests and ethical values making it possible for a partnership to be established on solid ground.
3. Identify common grounds of professional interest and mutual benefits. Look for complementation between both (or more) organizations, developing a common vision to gain institutional commitment from all sides. It is important not only that the heads of the partner organizations are committed to a relationship, but also those who will be responsible to make it work and to deliver the expected results.
4. For ease of creation establish the partnership with written agreements, clearly defining expectations from both sides. Include terms for confidentiality and transparency. The written

agreement should also include how copyrights and intellectual property rights for a joint product are going to be managed.

5. Develop a work program with an explicit description of the goals of the partnership. Define measurable targets and deliverables. Also specify clearly established roles, responsibilities, activities, timing and deadlines.
6. Explicitly state the responsibility of each partner regarding the contribution of financial, material, intellectual and human resources to the purpose of the partnership.
7. Draft a formal plan for measuring outcomes including mutually defined qualitative and quantitative metrics should be made explicit as well.
8. Communication is a priority, keeping a frequent an open dialogue to overcome misunderstandings and ensure longevity of the relationship. Get along with partners.
9. Agree on the terms of the formal termination of the partnership once the objective has been reached. Develop exit strategies to minimize costs if the partnership is not viable. Renegotiating the parameters of the partnership is easier if roles are clear from the start.

Although the term public institutions above refers to those organisations involved in the production of official statistics, the good practices for partnerships described before may also be applied to other public institutions and can be generalised to be applied more widely.

Results of Global Survey on Big Data 2015

The results of the survey indicate that Big Data sources of web-scraping (including social media), scanner data and satellite imagery are being used often, because they are generally “easy” to obtain and there is no cost associated with them. In the contrary, mobile phone data has generated a lot of interest, however, it’s not easily accessible (or requires complex customised agreement and lengthy negotiation with mobile phone operators) due to concerns of data protection and privacy issues. However, a partnership with mobile data operator is ranked high as the most data sources that countries are trying to establish/access.

These two separate challenges require two different solutions: more guidance (in term of methods and techniques) on accessing web-scraping, scanner data and satellite imagery and establishing global agreement with mobile phone operations that can transcend to national level. It should be noted that even though obtaining data through web-scraping is quite popular nowadays, it may change in the future, as many website operators/owners start to prohibit access through legal means (amending terms of use not to allow web-scraping) or technical means (blocking the access).

Top partnerships in Big Data projects, as identified by the survey, are as follows: government institutions, Big Data provider (e.g., retail chains, internet payment gateways, mobile data operator), research and academic institutes. It’s important to underline that statistical organizations need to start/keen establish good institutional arrangement with other government institutions that are probably collecting kind of Big Data that may be useful for the compilation of official statistics (as an example road sensors).

Questions related to the “identification of good practices of data access and partnerships”:

Does your office have a framework for dealing with privacy issues which also applies to Big Data sources? **Yes 60.3% No 39.7% (68 answers)**

Summary of comments:

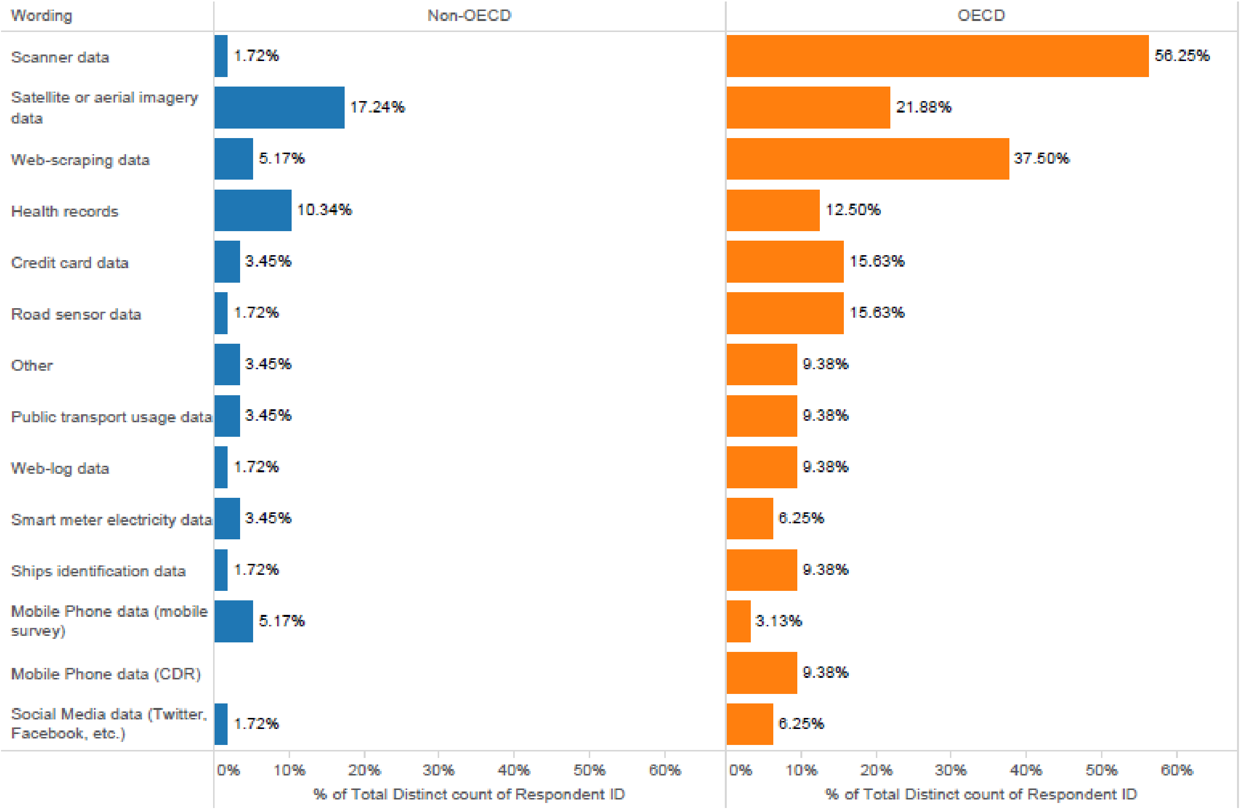
Majority of respondents explained that existing framework that is currently used to ensure individual data protection and to deal with privacy issues also applies to Big Data sources in addition to traditional sources (surveys and administrative data). However, those who answered No mentioned that statistical laws (on privacy) only applies to official data sources (data that is collected by government institutions), therefore amendment to existing regulations or establishing new ones are necessary.

Did your office address aspects of access to data, privacy and confidentiality specifically for Big Data? **Yes 28.6% No 71.4% (70 answers)**

Summary of comments:

Seventy-one percent of respondents said that there is no specific effort to address aspects of access to data, privacy and confidentiality. This in line with the answer of previous question regarding the use existing framework on privacy issues for Big Data sources (therefore no need for other legal framework). Others noted that privacy and confidentiality issues are addressed in the contracts signed with the data providers (ad-hoc approach). This includes several aspects such as anonymity data, access protocols, data transfer and microdata threshold for visualization. One respondent underlined that due to the heterogeneous nature of Big Data sources is difficult to solve via general measures (e.g., statistical act).

Which specific Big Data sources have you used or do you consider using?

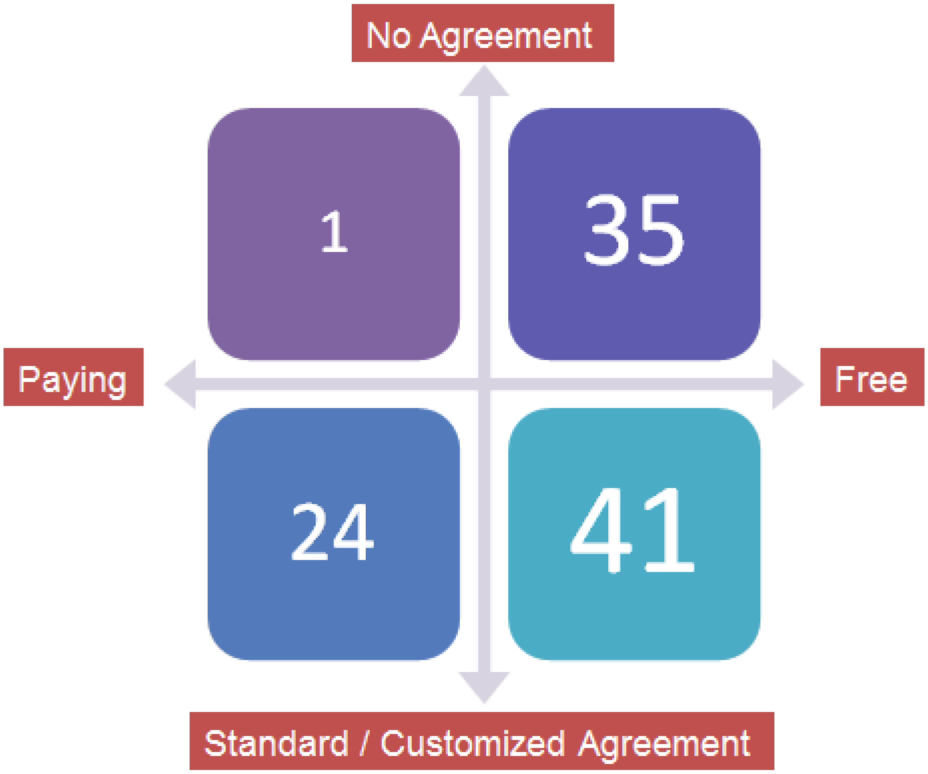


Summary of comments:

There is quite a significant variation between data sources in non-OECD countries and OECE countries. OECD countries identified the top data source that they have used and are considering using as scanner (price) data and web-scraping data (e.g., for price statistics), whereas non-OECD countries satellite or aerial imagery data, health record and mobile phone data.

How did you obtain access to the data? (part of project information, **total of 74 answers**)

	Free	Fee	Other
No written agreement necessary	29.7%	0.0%	
Standard agreement exists	17.6%	2.7%	
Customized agreement required	25.7%	8.1%	
Other			16.2%



Summary of comments:

From total of 74 responses, only 3.6% respondents indicated that access to the data involved some kind of a fee. The majority (73%) accesses the data for free, for examples: 1) publicly available data such as price information on the Internet (through web-scraping technique), 2) satellite imagery using standard agreement (through Creative Commons Licensing), or 3) scanner data using standard/customized agreement (through agreement with retail stores). Please note that for web-scraping, even though it is considered publicly available data, some respondents have established customized agreement with the respective website owners.

The cost components can be broken down into pre-processing cost, more detailed data (e.g., satellite imagery from commercial satellite), training of the staff, IT equipment/software and administrative cost.

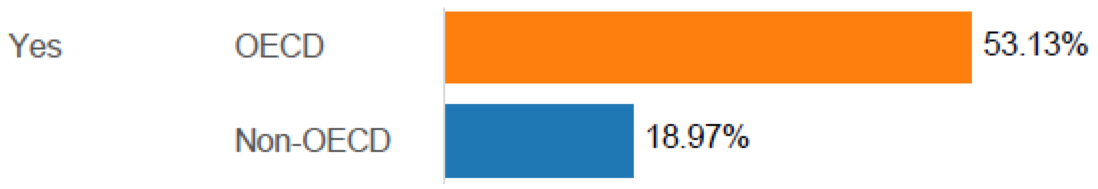
Many respondents indicated that they are still in exploration stage (including negotiation with data providers/owners), therefore there is no cost or the cost component is still unknown.

Do you use an intermediary (company or research institute) to obtain and prepare the data for your office?
Yes 17.9% No 82.1% (67 answers)

Summary of comments:

Most of the respondents (82.1%) did not use an intermediary to obtain and prepare the data. Those that used the intermediary including: pre-processing of raw satellite data by Geoscience Australia, scanner data from research institute, web-scraping results from company, etc.

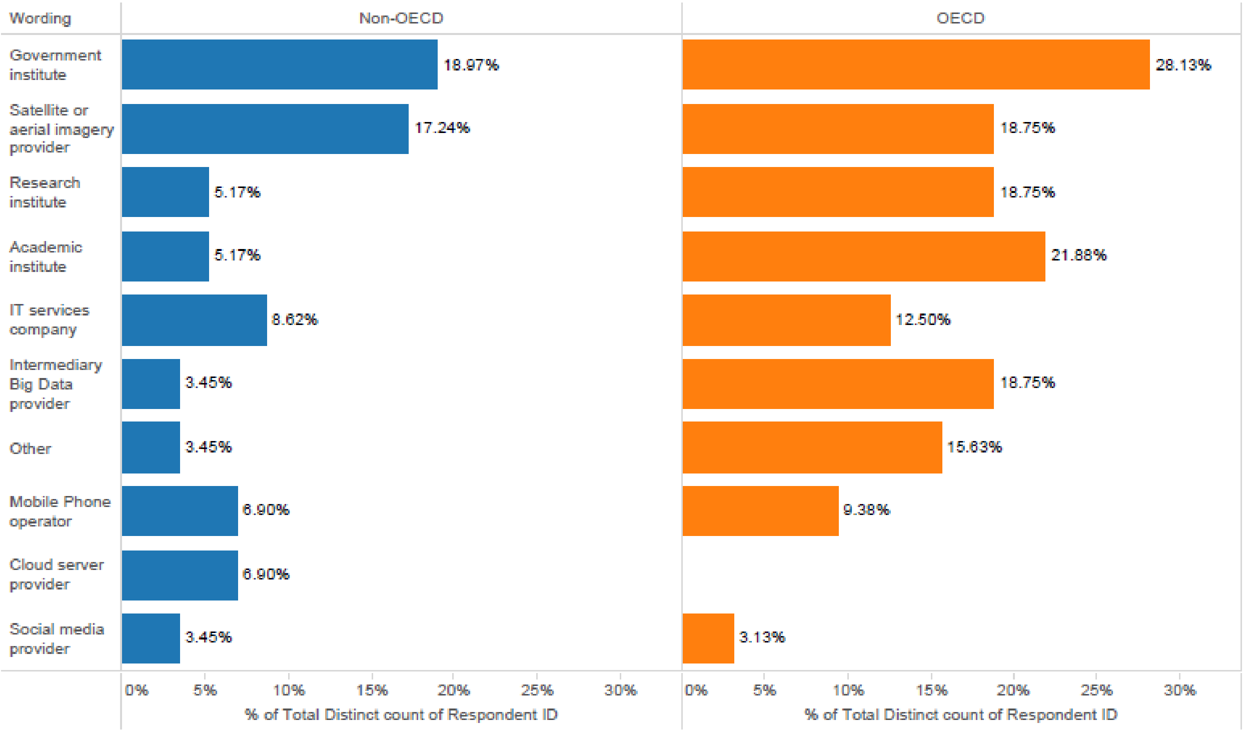
Have you established collaborations with other statistical agencies looking to use Big Data for statistical production?



Which Big Data partnerships have you established or are you trying to establish?

RANK	Has established	Trying to establish	Being considered
1	Government institute	Mobile Phone operator	Other
2	Satellite or aerial imagery provider	Government institute	Government institute
3	Research institute	Research institute	Mobile Phone operator
4	Academic institute	Academic institute	Academic institute
5	IT services company	Intermediary Big Data provider	Research institute
6	Intermediary Big Data provider	IT services company	Satellite or aerial imagery provider
7	Mobile Phone operator	Social media provider	IT services company
8	Other	Satellite or aerial imagery provider	Intermediary Big Data provider
9	Cloud server provider	Cloud server provider	Social media provider
10	Social media provider	Other	Cloud server provider

Broken down by OECD and non-OECD countries for established partnerships:



Summary of comments:

"Mobile phone operator" is the most popular partnership, but most countries are the level of "trying to establish/being considered". The second most popular partnership and also the success story is the "Government Institute" both in OECD and non-OECD countries. This elevates the relative importance of well-established cooperation among government institutions. In addition, "Satellite or aerial imagery provider" is cited as the second partnership that has been established so far. The other categories including financial institutions, retail chains, road sensors provider, internet payment gateways, and other industry players.

Who are your partners in the Big Data project?

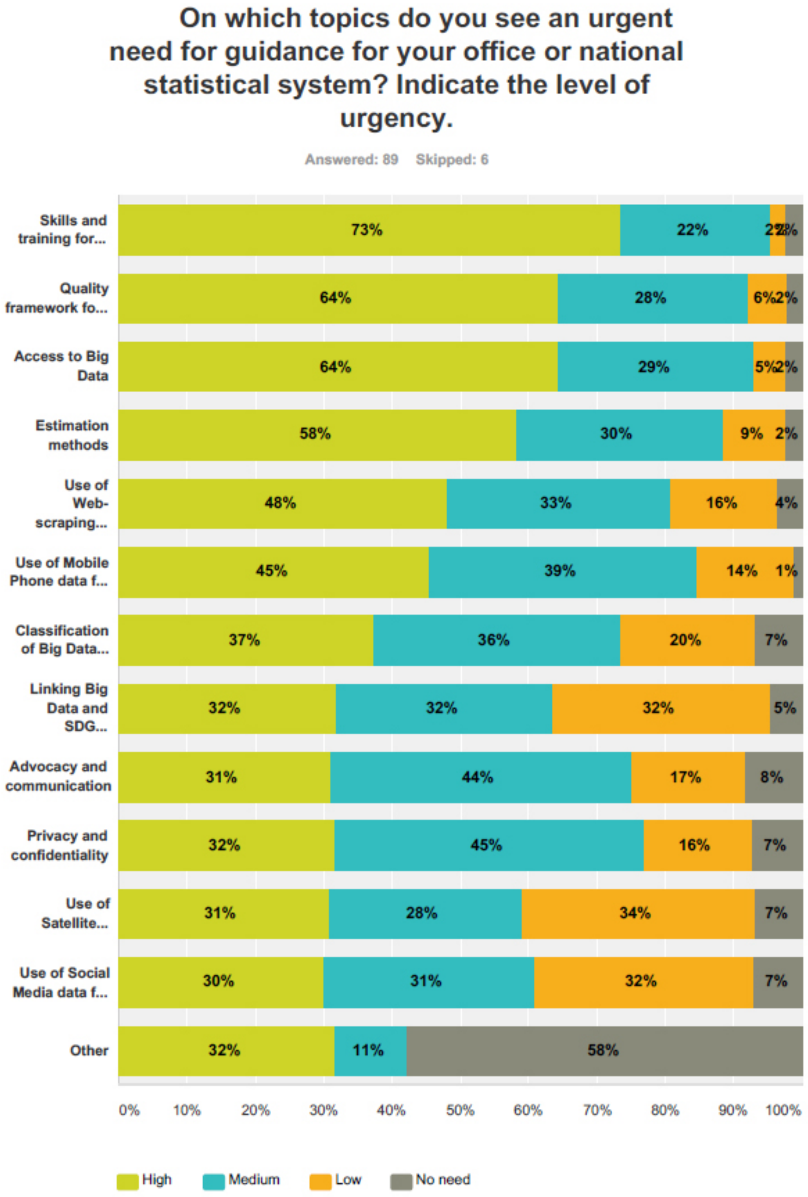
Partner(s):

Other	36
Government institute	34
Intermediary Big Data provider	16
Academic institute	16
Research institute	15
Mobile Phone operator	13
IT services company	10
Satellite or aerial imagery provider	5
Social media provider	3
Cloud server provider	3

Summary of comments:

“Government institute” as expected is the top partners in Big Data projects. It is followed by Intermediary Big Data provider, Research and Academic Institutes. “Other” partners constitute the largest chuck, they include the following: retail chains, news providers, internet payment gateways, and other internet website owners.

When asked about the need for urgent guidance NSOs clearly state the fields where advice is required. This is by itself a declaration on the importance is establishing partnerships on specific subject matters. On top comes the need for skills and training. The graph below from the Global Survey shows the priority themes for NSOs for which a working relationship is needed.



General conclusion/recommendation

Many countries have moved forward with using Big Data sources for official statistics, especially the legal framework on data access and privacy issues also applies to Big Data sources. In addition, it’s been also aided by “free” Big Data sources such as social media data, publicly available information in Internet (accessible through web scraping), satellite imagery and scanner data do not require complex data access agreement (or even require no agreement). On the other hand, accessing mobile phone data is quite challenging, as mobile phone operators are concerned with ensuring individual data protection and privacy issues, and possible data breach. Few options to resolve these issues are to address explicitly in the data access agreement and to involve actively mobile phone operators as partners in data processing/analysis. Nevertheless, NSO should establish good relationship with other government institutions as the source of Big Data, as they do increasingly collect various data that can be useful for official statistics.

It seems that there is a high correlation between availability of IT skills/tools in NSOs and the Big Data accessibility. This underlines the importance of possession of IT skills/tools in NSOs. In fact, in the survey, countries have identified that the skills and training for Big Data (e.g., the use of tools to facilitate access) and methodological issues (e.g., quality framework), are considered as two main issues to be addressed. Countries have indicated that “Access to Big Data” is considered one of urgent aspects of Big Data that require guidance. Access to Big Data, to IT skills, training and other important fields of knowledge such as linguistics, mathematical modelling and geospatial frameworks, as well as access to computing infrastructure to handle large data sets, can be better achieved by establishing partnerships with the appropriate parties.

It has been widely recognized that NSOs cannot alone address all the relevant issues pertaining to Big Data. Implementing a strategy to identify the needs for partnerships and for conducting and operating those partnerships is a relevant element in the overall project of using Big Data to complement official statistics and to improve the NSOs’ capacities for the future.

.