89383

Open Data Challenges and Opportunities for National Statistical Offices

Study Group Working Paper

1 July, 2014

This document reflects the results of a series of discussions and recommendations of a working group convened by the World Bank Group. The working group members are listed in Annex 1. The findings, interpretations, and conclusions are entirely those of the individual participants of the working group, and are intended to disseminate the findings of work in progress quickly and encourage an exchange of ideas. This is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. This working paper does not necessarily represent the views of the World Bank Group, its affiliated organizations, or organizations or governments with which the members of the working group may be affiliated.

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

© 2014 International Bank for Reconstruction and Development / The World Bank
1818 H Street NW
Washington DC 20433
Telephone: 202-473-1000
Internet: www.worldbank.org

The findings, interpretations, and conclusions expressed in this work do not necessarily reflect the views of The World Bank, its Board of Executive Directors, or the governments they represent.

The World Bank does not guarantee the accuracy of the data included in this work. The boundaries, colors, denominations, and other information shown on any map in this work do not imply any judgment on the part of The World Bank concerning the legal status of any territory or the endorsement or acceptance of such boundaries.

Rights and Permissions

For questions or comments concerning this working paper, please contact Timothy Herzog (therzog1@worldbank.org) or Amparo Ballivian (aballivian@worldbank.org).

Open Data Challenges and Opportunities for National Statistical Offices

Contents

1 Executive Summary

Open Data initiatives are transforming how governments and other public institutions interact and provide services to their constituents. They increase transparency and value to citizens, reduce inefficiencies and barriers to information, enable data-driven applications that improve public service delivery, and provide public data that can stimulate innovative business opportunities.

As the gatekeepers of official statistics, National Statistics Offices (NSOs) produce many datasets that could typically comprise the foundation of an Open Data program. They may also have relationships with other data producing agencies in the national statistical system and have expertise in dealing with the many technical and data quality issues attendant in publishing data. In short, they are extremely well placed to make a valuable contribution to Open Data initiatives.

Despite these advantages, NSOs do not always feature prominently in government-sponsored Open Data programs and they may be missing an important opportunity to expand the use and re-use of the data they produce. **The goal of this working paper is to better understand the opportunities and challenges that Open Data presents to NSOs and to identify what steps and solutions are needed to enable NSOs to play a valuable role in national or sub-national Open Data initiatives.** The working paper is the result of discussions among a study group whose members are listed in Appendix 1. The key findings are summarized in the next few paragraphs.

Open Data initiatives have significant implications for NSO operations

NSOs often manage many of the products, such as census records, economic statistics, and demographic data series, which are considered essential, high-value datasets in Open Data programs; hence, NSO products and expertise will be in high demand. The underlying principles of Open Data—to make government data as accessible and useful as possible—are clearly aligned with the NSOs core mission. Some requirements of Open Data are likely to be easily achievable by NSOs, such as opening data that is already public, and allowing data that is already available in non-open formats (such as reports) to be downloaded in bulk formats. Other desirable features of a good Open Data initiative, such as the transparent publishing of relevant policies, metadata, and training materials, will likely take more effort.

NSOs have strong roles to play in Open Data initiatives

While NSOs are not usually well positioned to lead a government-wide Open Data initiative, they nonetheless can be a vital component of it. NSOs produce high-demand official statistics, and can be instrumental in ensuring that the Open Data initiative is properly aligned with the wider National Statistical System. NSOs have extensive experience in data curation, the application of standards, and provision of metadata; hence, NSOs have a clear role to play in providing guidance to other agencies in publishing their own data. Finally, as members of the international statistics community, NSOs are in a good position to make sure that the Open Data initiative is well aligned with the efforts of other countries and international organizations, and follows internationally agreed standards.

Open Data initiatives can greatly benefit NSOs

Increasing the use and applicability of high-quality data not only goes to the heart of the Open Data agenda, it is the basis for why NSOs exist. Data collection is a costly activity, financed by public spending and other investments. Open Data can increase the return on those investments by expanding the ways that NSO data are used by a variety of users. Open Data can also relieve pressure on NSO operating budgets by reducing the demand for custom tabulations and other data requests. Open Data will likely raise the profile of NSOs both within the government, with key constituencies, and with the public at large because of the unique roles the NSO can play. A greater profile can in turn enhance the NSOs reputation and open additional opportunities.

Data quality is often cited as a concern preventing NSOs from adopting Open Data policies. Almost all statistics are subject to various quality factors which are intrinsic to standard statistical methods. The important thing for NSOs is to provide adequate context and information about the uncertainties and limitations of any particular dataset (open or otherwise). Doing so would lead to increased transparency and greater public trust in both the data and the NSO itself. Furthermore, NSOs can use feedback mechanisms available through Open Data to improve data quality.

Open Data raises potential challenges for NSOs

NSOs have the responsibility, often established in legislation, to protect the confidentiality and privacy of their data providers, who may be individuals, households or businesses. Confidentiality issues do not apply equally to all types of data, and many types of data, such as aggregated statistics, can typically be published and opened without breaching confidentiality so long as standard anonymization techniques are applied. In the more sensitive realm of survey microdata, each NSO must ultimately make the determination about whether and how to make these data public and which anonymization techniques to use, but experience to date includes several cases where microdata have been opened without compromising confidentiality.

Open Data best practices require that data producers provide clear terms of use with minimal restrictions on how data can be used. This may prove challenging for NSOs that manage different classifications of products, some of which may be restricted on grounds of confidentiality or even national security. However, NSOs can take advantage of standard international licenses, and there are several case studies for managing data under multiple access policies.

NSOs may also be concerned about the resources and capacity required to implement Open Data. Experience to date suggests that the additional resources are not substantial, and may be at least partly offset by cost savings and greater efficiencies. But it is true that Open Data may represent an opportunity costs to some NSOs, to the extent that they derive some revenue from data sales.

NSOs may also face challenges in engaging a larger user ecosystem. Since the goal of Open Data is to increase the use of data, it is almost inevitable that the NSO will be interacting with user communities which are new and possibly unanticipated. Again, this can be an opportunity for the NSO as much as a challenge. Many public agencies that have adopted Open Data policies have experimented using workshops, toolkits, competitions, social media and other approaches to successfully engage new data users. Ultimately, what may be required is for NSOs to simply be open to change, in order to reap the benefits that Open Data can provide them.

2 Introduction

Open data initiatives are transforming how governments and other public institutions interact and provide services to their constituents. Since 2009, over 43 countries—and many more states, provinces and cities—have launched Open Data initiatives. Governments are using Open Data to deliver increased transparency and value to citizens, reduce inefficiencies and barriers to information, and provide public data that can stimulate innovation and economic opportunity.

However, National Statistical Offices (NSOs) do not always feature prominently in government-sponsored Open Data initiatives, particularly in developing countries. As the agencies responsible for maintaining and disseminating a country's official statistics, NSOs are the "gatekeepers" for many of the datasets that typically could comprise the foundation of an Open Data program. NSOs may also have relationships with other agencies that provide raw data to the statistical system, endowing them with an important role in the local data community. Finally, NSOs typically have expertise in dealing with the many technical and data quality issues attendant in publishing public datasets, making them valuable knowledge resources.

Despite these advantages, NSOs may be missing important opportunities to expand the use and re-use of the data they produce, increase the return on investment in statistical production, and play a more prominent role as leaders of national statistical systems.

The goal of this working paper is to better understand the opportunities and challenges that Open Data presents to NSOs and to identify what steps and solutions are needed to enable NSOs to play a valuable role in national or sub-national Open Data initiatives. The working paper summarizes the discussions and recommendations of a working group of experts from the NSO and Open Data communities, to explore Open Data from the perspective of NSOs.[1] The working paper is intended for:

- **NSO senior management and staff.** This working paper provides guidance for NSOs on implementing Open Data initiatives that are linked to national statistical systems, and includes information on a range of technical issues.
- **Other government agencies and ministries.** Particularly for agencies charged with leading a country's Open Data initiative, this working paper highlights how the leaders of a national or sub-national Open Data initiatives can use their NSOs to provide valuable insights and expertise to the initiative and to other collaborating agencies.
- **International agencies.** Organizations that work to strengthen and support national statistical systems and NSOs in developing countries—including the World Bank, the U.N. Statistical Division, other U.N. agencies, regional development banks, and PARIS21—will find insights on how Open Data can be advantageous to NSOs.

[1] The composition of the working group, convened by the World Bank, can be found in Appendix 1

3 What is Open Data?

"Open Data" is generally understood to mean data that are made available to the public free of charge, without registration or restrictive licenses, for any purpose whatsoever (including commercial purposes), in electronic, machine-readable formats that are easy to find, download and use. As applied to public institutions such as governments and intergovernmental organizations, Open Data is grounded in the recognition that government data is produced with public funds so, with few exceptions, should be treated as public goods.

In addition, Open Data can produce large social and economic benefits[2]. A key characteristic of Open Data is the potential for reusability, both by data experts and the public at large. Reusability is the key to creating new opportunities and benefits from government data as detailed later in this working paper. For Open Data to be reusable it must generally meet two basic criteria. First, the data must be *legally* open, meaning that it is placed in the public domain or under liberal terms of use with minimal restrictions. This ensures that government policies do not create barriers or ambiguities concerning how the data may be used. Second, the data must be *technically* open, meaning that it is published in electronic formats that are machine-readable and preferably non-proprietary (see "5 Stars of Open Data" in Appendix 2). This ensures that ordinary citizens can access and utilize the data with little or no cost using common software tools.

Data that are publicly available do not necessarily meet the definition of "Open Data." For example, data from NSOs may be publicly available, but only to certain qualified or registered users, or with narrow restrictions on how the data can be used. Data may also be publicly available but only in proprietary formats that are difficult to access or manipulate (such as PDF), or even non-electronic formats. Nonetheless, these data are often characterized as "public data." The advent of "Open Data" standards has the potential to create confusion if users do not understand the difference. It is therefore important for NSOs (and other organizations that distribute data under multiple policies) to clearly differentiate their data products and corresponding policies. This issue is further explored under "Addressing Privacy and Confidentiality" as well as "Legal, Licensing and Policy Questions."

[2] Some of the most developed countries in the world have recognized this. See, for example, the G8 Open Data Charter: http://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex. A recent McKinsey report estimates the potential economic benefits of Open Data to the United States economy to be around 3 trillion U.S. dollars per year: http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information

4 Significance of Open Data to NSOs

There are several reasons why the Open Data movement is of particular significance to NSOs, and why it is important for an NSO to be actively involved in its government's Open Data initiative.

First, NSOs typically manage many of the data products considered essential or at least high value in government-wide Open Data initiatives. This means that while the Open Data initiatives vary in focus among countries, data managed by NSOs is frequently in high demand. Even in countries that have decentralized statistical systems, NSOs may play a role in coordinating data from these systems, and thus have a clear opportunity and role to play in the context of Open Data.

Second, the underlying principles of Open Data are clearly linked to one of the NSOs' fundamental purposes: to make relevant statistics available in ways that are easy for users to access and use. NSOs are perhaps unique with respect to other government agencies in that production of statistics is the primary function of the NSO, as opposed to administrative, budget or other types of data which might be considered "byproducts" of government functions. The principles of Open Data are thus especially relevant to NSOs. Releasing outputs in open formats can support this task without being significantly disruptive or costly. For instance, if an NSO already publishes aggregated statistical outputs in periodic reports or similar forms, it could simply publish the same outputs online using Open Data formats and attaching an open license. Thus the NSO would quickly achieve a significant improvement in Open Data implementation. This would quickly put the NSO at the 3 star level in the 5 star model of Open Data (see Appendix 2 for an adaptation of the standard 5 star model to NSO data). At this level, data are free of charge, available for re-use without restrictions, indexed and easily discoverable and in non-proprietary machine-readable formats.

Open Data can be strategically important to NSOs in several other respects. Broader use of NSO-supplied data, throughout both the government and among citizens, can create pathways to increase efficiency of dissemination, improve data quality, lead to the modernization of administrative records, and raise the public profile of the NSO. These opportunities are further explored later in "Challenges and Opportunities of Open Data for NSOs."

Some government-produced data cannot be published nor opened. Open Data policies typically allow for reasonable exceptions to be made for national security and to safeguard the privacy of individuals, households, and businesses. However, even with these few exceptions, there are still vast opportunities for many categories of government data at many levels to be opened to the public. Many NSO data outputs will be straightforward to release as Open Data, including:

- Statistical products that are already publicly available without restriction—perhaps through printed publications, the NSO's website, or upon request
- Other vital census and economic statistics at the national and sub-national levels
- Price and trade data
- Registers used for drawing statistical samples, for example lists of businesses

- Official maps of political boundaries, voting districts, infrastructure, and the location of public facilities (schools, government offices, police stations, libraries, etc.)
- Classification systems such as for household consumption or types of industry

MetadataFor context, examples of information that could be released as Open Data from other branches and levels of government might include:

- Budget and expenditure data at national and ministerial levels
- Procurement, contract award and grant records
- Parliamentary records of proceedings, draft legislation, and enacted laws
- Public service delivery and performance data, for instance for schools or health clinics
- Transport data including roads and public transit
- Weather, environment and agriculture data such as forecasts, air and soil quality
- Real estate, construction, zoning and property records
- Labor statistics and employment data
- Investment and trade data

The types of NSO data listed above do not present particular privacy issues and are not sensitive once released. In general, there should be no barrier to releasing these outputs in open formats and NSOs should be supported and encouraged to do this quickly and efficiently.

Other statistical products present challenges in terms of dissemination. Many micro datasets fall into this category, since the NSO must maintain the confidentiality of individual respondents (individuals, households, employees or businesses) in the survey sample. For this reason, it is important to recognize that NSOs will likely not employ a "one size fits all" data policy; more likely they will employ a range of dissemination policies depending on the needs and sensitivities of each dataset, and in accordance with legal requirements.

5 The Role of NSOs in a National Open Data Program

NSOs have many important roles in a national—or even sub-national or sectoral—Open Data program, including:

- Contributing data to the national Open Data initiative. As previously noted, NSOs often produce the majority of a country's official statistics and so they will be significant contributor to the supply of data.
- An NSO may develop its own Open Data platform, or one that covers the wider national statistics system. In some countries, the NSO's data appear both on its own portal and the government-wide platform.
- Given the likely higher experience of NSOs in techniques and standards related to data collection, curation and release—for example metadata and data anonymization—they will have a clear role to play in issuing guidelines in these

areas for their own statistical products and for other agencies in the national statistics system. This topic is further developed below in section 6c.

- NSOs can also have a training and leadership role for these statistical issues within a national Open Data program.
- NSOs typically have strong links with statistics offices in other countries and statistics departments in international agencies, such as the U.N. Statistical Commission, the World Bank and the regional development banks. Collectively these NSOs and international agencies have developed international standards for the collection, processing and dissemination of data. These relations thus provide a vehicle for the development of Open Data standards for the international statistical community. In addition, the international community can often be more effective in disseminating data than individual NSOs would be on their own.

It is also important to clarify what roles NSOs are *not* expected to play. There will be many different ministries and agencies involved in a country's Open Data initiative. In general, national Open Data policies are most successful when they are led by the Office of the President or Prime Minister, or by a central government ministry such as planning or economy. This reflects the wide and crosscutting nature of Open Data initiatives and the need to have leadership from a ministry that is able to coordinate all relevant parts of government. NSOs will have a clear role to play, but they are not expected to take an overall lead, nor is it appropriate for them to do this. In the worldwide experience to date, the wider coordination, legislative and licensing issues related to national Open Data initiatives are outside the mandate of the NSO.

6 Challenges and Opportunities of Open Data for NSOs

Making the move from publishing data to releasing statistics as Open Data is a significant step for an NSO. It is likely to require a culture change. As such it is natural to expect that it will present certain challenges. However, NSOs can use the process to redefine their relationship with citizens, government and other users to become more outward-focused and open. This section goes through the most significant of these challenges to outline the background to each of them and to provide information on how they can be approached and dealt with.

In addition to the challenges, however, Open Data also has the potential to deliver significant benefits and opportunities to NSOs. We start this section by discussing these benefits.

6a Benefits of Open Data for NSOs

6a(i) Increasing the use and re-use of data

Getting more, better and wider use and re-use of data goes to the heart of the Open Data agenda. This is one of the key benefits to NSOs of making their data more open.

Data collection is not a cheap activity. And almost always the funding for data collection within a government's national statistics system comes from the domestic taxpayer and, for countries that receive aid, from taxpayers in donor countries. It is therefore important to deliver as much value as possible from the resources that go to government statistics.

At present NSOs typically release a range of outputs in a relatively fixed series of formats. Further detail, additional analyses and alternative formats are sometimes available on request or for an additional fee – but these processes can easily act as a barrier to many potential users.

Releasing existing NSO outputs and additional underlying data as Open Data will lead to a wider range of users being able to access, use, manipulate and combine NSO data in new and innovative ways. If data are put to only one use, the return to investment in data production is given by that single use. But if the same data are used and re-used several times, this undoubtedly increases the return to investment in data production.

6a(ii) Reducing costs of data dissemination

Making data available as Open Data allows users to access the information directly without incurring NSO staff costs and other costs, for example associated with custom tabulations. In addition, NSOs are often required to send data to international agencies in specified formats—requirements that are both time-consuming and costly. Making data available as Open Data can reduce both time and costs related to international reporting. Some NSOs that have already moved to making their information available in open formats have seen direct requests for data decrease.[3] This enables scarce staff resources to be re-directed to other tasks.

6a(iii) Raising the profile and influence of the NSO

Participation in Open Data initiatives can have great potential to increase the profile and influence of an NSO for two main reasons. First, NSOs will likely have a role in providing technical assistance to other data suppliers to the national Open Data portal (as discussed in Section 5). Second, a national Open Data portal is likely to have a different audience than the NSO website including new types of users; therefore NSOs can increase their data dissemination at no additional cost. A broader audience can in turn lead to additional uses of NSO data; recognition of the importance of statistics and therefore the need to invest resources in it; and benefits in terms of the ability to direct and deliver a coordinated and coherent national statistics system.

In the past NSOs, particularly in developing countries, have typically had a relatively low profile. They have produced a relatively fixed range of outputs for a fairly limited range of users. However, in recent years, the debates around evidence-based policy-making and aid effectiveness have led to a renewed focus on the importance of investing in statistics and producing high quality and relevant data. The work of PARIS 21 through

[3] There will still be room for, and the need for, the possibility of users requesting custom tabulations. But releasing statistics as Open Data can reduce the resources required for this.

encouraging National Strategies for the Development of Statistics (NSDSs) has been particularly instrumental in this area. The participation of NSOs in Open Data processes can build on this work to increase the profile and influence of both NSOs and the statistics they produce. NSOs may even be able to negotiate successfully for additional budgetary resources to implement Open Data in light of this fact.

Open Data initiatives naturally bring together a wide range of government ministries and agencies to work together. Again, NSOs can potentially be significant players in this due to the large amount of data that they can contribute and their specialized knowledge concerning the preparation and description of data for release. This can provide the opportunity for a stronger role within the public sector, the chance to promote use of data in evidenced-based decision making both within and outside the public sector, and help maintain strong and wide support for the continued investment in statistics. NSOs can also help establish norms and standards for open government data. And they will be able to guide and influence other parts of government through their leadership in the wider national statistics system.

In short, Open Data initiatives can enable NSOs to increase their relevance and engagement with the wider work of government. In some instances, NSO efforts on Open Data have even received the attention and accolades of popular media. For instance, the U.S. Census Bureau was recognized in 2013 for its Open Data API[4].

6a(iv) Improving data quality

The issue of data quality is often mentioned during discussions with NSOs about Open Data. In particular, poor data quality is sometimes given as a reason for a reluctance to release statistics in open formats. NSOs should, of course, be able to draw the line between what should and should not be released from a quality perspective, and datasets that have not finished quality control processes should not be published prematurely. Managed correctly, however, greater openness can actually help NSOs address data quality issues – and therefore it is a reason for embracing Open Data rather than avoiding it. Open Data will certainly introduce new issues to consider in this area and may require changes in practices and institutional thinking to deal with situations appropriately.

Almost all statistics are subject to various quality issues, for example sampling and non-sampling errors, preliminary estimates that are later revised, and other data collection and measurement issues. In terms of data quality, the key issues are whether the statistics are fit for the purpose for which they are released and whether data producers are transparent about the characteristics of the data. NSOs routinely release data with varying degrees of uncertainty, coverage and other characteristics, and sometimes information about what conclusions can and cannot be made. In fact, when an NSO is transparent about data quality issues, its candor can actually increase the public trust in the data and the NSO itself. Many data users are well used to dealing with imperfect data and the caveats that go alongside them.

[4] http://www.census.gov/newsroom/releases/archives/miscellaneous/cb13-tps37.html

Particular quality issues will vary depending on the type of data being considered for open release. For some types of data, the issues will be minimal or non-existent. This applies, for example, to aggregate statistics and indicators that have already been publicly released (although not necessarily in open formats). There should be no barrier to releasing these outputs as Open Data from a data quality perspective as they will already have been put through the NSO's quality assurance procedures.

In other situations, more disaggregated data may be considered for release as Open Data. These data may not yet have been publicly released and the NSO may feel it has not been subjected to such stringent quality control processes. However if the data are deemed to be of sufficient quality to feed into public policy making (through the aggregated statistics) then it follows they should be of sufficient quality for the wider general public. And if there are known quality issues, then those issues should be documented, and there should be processes in place—including timetables—to outline how and when they will be improved. In other words, quality assurance should not be an excuse for restricting data availability.

As noted above, wider and more intense use of data can often help NSOs improve the quality of their data. The more statistics are compared, contrasted and combined with other data and information, the more light is shed on quality issues that may not have been identified previously. More "eyes" looking at data increases the likelihood of identifying inconsistencies, gaps and other quality issues. In this sense, users can play a role as quality reviewers—free of charge. Furthermore, releasing statistics as Open Data has the potential to allow users to explore data in new and innovative ways. Inevitably this will lead to greater feedback which, used appropriately, can help NSOs improve their outputs.

To take advantage of the opportunities afforded by Open Data for data quality improvement, this working paper makes the following recommendations:

- Publish the data quality guidelines used by the NSO (and guidelines used by other agencies), along with the publishing schedule for each dataset, so that users know when data and data revisions will be available and can hold data producers to account.
- Do not wait for perfection before releasing data as Open Data; rather embrace Open Data in order to use it as a pathway to improve data quality.
- Actively use Open Data to get feedback from users on quality issues relating to the statistics. For example, links could be provided from the Open Data portals to allow users to provide comments to the NSO or other data producer.
- Be transparent about methods, strengths and limitations – and provide information to users on how the data should and should not be used. Metadata will be a key component in this area, describing, for example, the data collection processes, the format and definition of the variables, the purpose of the data and other key characteristics.
- Use the release of data to help the NSO think through what improvements could be made, and how and when these will be done.

6b Challenges of Open Data for NSOs

6b(i) Privacy and confidentiality

The issue of maintaining the confidentiality of data providers is frequently raised as a barrier to Open Data. Ensuring confidentiality is a concern for many types of government data, and the responsibility to maintain confidentiality applies with or without the presence of an Open Data initiative.

For aggregate statistics and indicators, Open Data generally does not present additional confidentiality issues beyond those for traditional release methods[5]. The issue of confidentiality applies particularly to some microdata[6]. When these data are obtained from individuals, households or businesses, they provide their information to NSOs on the assurance that their privacy rights will be respected and they will not be identified in the data or results that are released. In many countries, the requirement to protect a respondent's confidentiality is based in statistical legislation. Furthermore, ensuring confidentiality is essential to safeguarding respondents' trust, obtaining high response rates and accurate responses.

Experience thus far indicates that NSOs have used a variety of approaches to making microdata publicly available. In some cases, the Open Data release of microdata may be relatively straightforward from a confidentiality perspective. Examples of this include the release of individual price data from the CPI[7], heavily anonymized survey microdata (so-called "training datasets"), and synthetic microdata. At the other extreme, microdata access is highly restricted, available only to qualified users for sanctioned purposes, such that microdata are not considered part of the NSOs Open Data products.

Large sample sizes and use of anonymization techniques may allow the NSO to release many (but not all) of its microdata products under open terms of use. Appendix 3 includes a summary of several microdata dissemination programs. The conclusion of this working paper is that the unique circumstances of each microdata product and the professional judgment of each NSO should determine whether and how to publish its microdata, balancing the goals of openness and safeguarding of confidentiality, but that the objective should be to make microdata as open as possible.

The specific facts of each case will determine what approaches can be used to make a microdata product "safe" for release, and how wide and open this release can be. NSOs

[5] In rare circumstances aggregate statistics may present the possibility of identifying respondents, for example by taking the difference between two published figures which differ by only one respondent, or where a particular business dominates the market share of a product. In these cases, anonymization processes would be conducted even for the routine release of these statistics. These issues are generally no different or greater within an Open Data program than without.

[6] For the purpose of this working paper, microdata is a dataset that includes responses at the level of the individual observation. These respondents may be people, households, businesses or other units. The data for these response units may have been obtained from censuses, sample surveys or through administrative procedures.

[7] http://www.ons.gov.uk/ons/guide-method/user-guidance/prices/cpi-and-rpi/responding-to-the-open-data-agenda---an-ons-case-study--consumer-prices-index.pdf

have a range of anonymization techniques at their disposal. These techniques include the removal of all directly identifying variables, removing or substituting geographic identifiers, suppression of unique or unusual responses (e.g., top-coding or bottom-coding), and the introduction of small random errors or noise to some cases.

Restrictions on release may still be required for some sets of microdata, even after anonymization. These restrictions will depend on the sensitivity of the data. Ultimately the openness of release will come down to a determination from the NSO, but this should be done as objectively and transparently as possible. This implies that NSOs need to adopt privacy and confidentiality policies for data openness and publication, both for their own statistical production and for that of other public agencies.

The recommended approach is:

- Make microdata as widely available as possible, while recognizing that some types of microdata should not be published nor opened.
- Develop specific criteria to determine in an objective and transparent way which microdata will be released and how open and wide that release will be.
- If NSOs make microdata publicly available, but not under "open" terms of use (for example, only to registered users and/or for sanctioned purposes), the NSO should make clear distinctions between its "Open Data" products and its restricted products. For instance, restricted microdata products should not be registered in the NSO's or the government's Open Data catalog.

6b(ii) Legal, licensing and policy questions

Successful OGD initiatives require clear and consistent supporting policy/legal frameworks. For government-wide programs, best practices suggest a single license that is applied to all Open Data produced by any government ministry. Such licenses are very straight-forward and easy to read, and typically provide a simple, bulleted list of (broad) rights and (minimal) responsibilities, with links to longer supporting legal documents as necessary. Many open licenses, such as those of the United Kingdom[8], United States, [9] and France[10], the World Bank[11] and many others[12], are explicitly modeled after and compatible with one of the Creative Commons licenses [13] (usually either CC-By or CC-0). Open licenses may also reference underpinning legislation or principles in the Constitution, as is the case for Kenya[14]. Once a standard license is promulgated, NSOs or any other agency can "apply" the license to a dataset simply by referencing and linking to the license in the dataset's metadata.

[8] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/

[9] https://www.data.gov/data-policy

[10] http://www.scribd.com/doc/69411439/French-PSI-Re-Use-Licence-Licence-Ouverte-Open-Licence-ENG

[11] http://data.worldbank.org/summary-terms-of-use

[12] http://data.worldbank.org/ogd

[13] http://creativecommons.org/licenses

[14] https://opendata.go.ke/terms-of-service

Along with promulgating user licenses, Open Data policies may also provide terms for data providers, i.e., government agencies and ministries, to supply datasets to the Open Data program. These policies are also public and may even appear on the same page as the user license, as is the case for Kenya[15] or the United States[16]. For suppliers, data policies provide clarity on issues such as proper citation, protection of privacy, data quality, exclusions for national security or other reasons, and metadata. Including this guidance in one place is helpful not only for agencies, but for users to understand the expectations (and limitations) of data providers.

Clear, legal guidance is especially important for agencies that manage data under two or more access policies. This is often the case with NSOs who manage micro dataset catalogs that are made available under one or more licenses that restrict how the data may be used, in order to protect the confidentiality of the survey samples. A simple, uniform, highly visible Open Data license can be easily applied to datasets such as vital records, national accounts, economic statistics, and any other data that the government considers open and for which confidentiality is not an issue. Conversely, microdata catalogs which allow users to search metadata (but restrict access to the actual data) can display the appropriate licenses in an equally visible fashion. A good example of this practice is the NADA Microdata Cataloging Tool[17], an open source platform from the International Household Survey Network (IHSN).

Complications can arise where Open Data policies do not align well with pre-existing legislation or institutionalized practices within various agencies. For instance, legislation intended to protect individual privacy may be interpreted (rightly or not) as preventing the release of aggregated national statistics that would normally be considered open. Legislation designed to protect state secrets or national security may forbid the release of certain data. In this case, the criteria for classifying data as closed for national security reasons should be public and as objective as possible.

6b(iii) Implementation and resources

As noted in section 6a(i) above, Open Data initiatives provide NSOs with opportunities for greater efficiency and cost savings in several areas. However it is also important to recognize that Open Data will also have its own set of costs. These must be acknowledged and addressed, but they are likely to be relatively small compared to other NSO costs, such as for conducting surveys, censuses and other data collections.

At the start of an Open Data program, an NSO will need to invest a certain amount of time and resources to familiarize staff with the principles and practices of Open Data. An NSO will also need a certain amount of new skills and experience within its office, for example preparation of data and metadata for release as Open Data, and engaging with new types of users (see the section below on user engagement). This implies either additional training for existing staff or the recruitment of new staff or both.

[15] https://opendata.go.ke/terms-of-service
[16] http://www.data.gov/data-policy
[17] http://www.ihsn.org/home/software/nada

It may also be worthwhile to consider changes to organizational structures within an NSO. Typically, dissemination activities in developing country NSOs have been organized around separate outputs (with separate teams) rather than as a single effort. Work on Open Data will go beyond traditional dissemination, but it will be worth considering whether a dedicated team within the NSO might be set up to handle Open Data activities working alongside each of the output teams.

There will be some costs associated with the new software and platforms associated with Open Data, but they typically are not prohibitive; startup costs range from $5,000 - $75,000 depending on the choice of platform, service level, and other considerations[18]. Open source software tools may provide cost savings since they are free to purchase, but only when the NSO has the necessary technical capacity to install and manage them. In other cases, commercial software may provide more economical alternatives. In some cases an NSO may use only the platform developed for a government-wide Open Data initiative, in which case the additional costs may not fall directly on the NSO budget.

As well as these additional costs, there may be resource implications from revenue foregone, in particular, when NSOs have a policy of charging for client-specific requests for datasets. In many cases NSOs only charge the specific costs for the additional work. Moving to Open Data will allow users to obtain data without going through the NSO, so the loss of revenue should be balanced by the additional staff time that is subsequently available. In a small number of cases an NSO may also charge a profit element when responding to requests for bespoke data products. In these cases Open Data will imply a loss of revenue, and discussions will need to be held with the funding ministries and partners to decide how to address this. However, in general, the benefits of the wider use of data resulting from Open Data will far outweigh these losses of revenue.

This working paper makes the following recommendations in this area:

- Plan how the move to Open Data will be made within the NSO, using advocacy materials and training for staff.
- Identify the additional skills and experience that will be required within the NSO; select staff that can be trained for these roles and recruit additional staff to fill remaining gaps.
- Work with the ministry leading the government's Open Data initiative, if there is one, to select and procure the appropriate software and platforms.
- Make an assessment of potential revenue loss and cost savings from moving away from charging for data, and use this information in discussions on resources with funding ministries and the leaders of the national Open Data initiative.

6c Integrating Open Data with national statistical systems, NSDSs and national M&E systems

Although the overall leadership of a government's Open Data initiative will usually lie elsewhere, NSOs will have a natural leadership role to play concerning Open Data within

[18] Estimates based on an informal survey of leading Open Data platforms.

the national statistics system. This has already been touched on above in the discussion on the role of NSOs in Open Data. This section continues this and focuses particularly on strategies to strengthen both national statistics and national monitoring and evaluation (M&E) systems.

NSOs are not the sole producers of government statistics; there are other ministries and government agencies that produce data. Together, the NSO and these other government agencies that produce data make up the supply side of a country's national statistical system. In many cases the NSO has a formal role to coordinate this wider system based on statistics legislation, while in other countries their role is more informal.

This concept of a wider national statistics system can be important in the context of an Open Data initiative. The NSO has the potential to provide guidance, standards, regulations, technical assistance and training to help other agencies move towards releasing their data in open formats. It can also provide its knowledge of statistics legislation to determine its consistency with Open Data licenses.

NSOs typically have mandates to develop strategies for their countries' government statistics. Since the approval of the Marrakesh Action Plan[19] in 2004, the international community has encouraged and supported NSOs in developing countries to produce a National Strategy for the Development of Statistics (NSDS) to set out how a country's statistics will be strengthened. PARIS 21 provides guidance on producing NSDSs and the latest revision includes a section on how NSOs can engage with Open Data[20].

Moves towards Open Data could also transform the ability to monitor and evaluate government activities and policies. If the institutions responsible for reporting on their activities and outputs were to make this information available regularly and in open formats, then a wide range of institutions would easily be able to take part in monitoring and evaluating the policies and programs rather than simply relying on a published M&E report.

This working paper makes the following recommendations in this area:

- Use the NSO's leadership role in the national statistics system to support other government statistics producers to engage with Open Data, as well as promote data best practices and techniques throughout government ministries.
- Review statistical legislation to ensure it is able to support moves towards Open Data.
- Incorporate Open Data into current and forthcoming NSDSs.
- Encourage ministries, agencies and other organizations involved in government activities to release data relevant to M&E as Open Data, and encourage a wide range of stakeholders to engage in the M&E of government activities through using this data.

[19] http://go.worldbank.org/PRTR3BCNE0
[20] http://nsdsguidelines.paris21.org/node/530

6d The Open Data ecosystem

Making NSO data open gives users the freedom to do what they want with it. However, there is also a need for data to be released that is relevant to users and in a form that makes it usable. This will require engagement with users to explore what data should be made available, including the metadata that accompanies it, and how this should be done.

Most NSOs will have existing methods of engaging with users. User/Producer groups are often established within national statistics systems to enable discussions between NSOs, other data producing parts of government and the main users, particularly in the process of developing the NSDS. However these groups are often limited in their membership, typically covering key government users, development partners and research institutes. In addition to these traditional users, the move towards Open Data is likely to result in a growth in the number and type of users; many of these, such as application developers, data scientists, journalists and advocates, may be unfamiliar to NSOs. This means that new methods of identifying, contacting and discussing with users will be needed.

As the Open Data movement grows, new methods and strategies to encourage the use of data have emerged[21]. One particularly successful strategy has been the use of competitions to challenge users to develop innovative uses for newly released data. Examples include creating new ways of visualizing data, combining and merging different datasets, and developing applications that add value to data and make it accessible. Events such as hackathons and codefests are common within the wider Open Data community[22] and NSOs should consider opportunities for involvement in existing and forthcoming events, as well as setting up their own.

Engagement with users and an increase in the use of data will also be encouraged if NSOs provide:

- **Metadata.** Metadata provides information about the data including how and when it was collected, definitions of variables, quality issues relating to the data, etc. Robust, accessible metadata is perhaps the simplest catalyst for increasing user engagement, and also one of the most effective.

- **Tools and guidance.** Use of Open Data can be increased by the development of tools and guidance related to Open Data or to particular datasets. One of the easiest ways that NSOs can provide insights and guidance is to use blogs to explain the characteristics of various products. For example, the World Bank uses its blog to regularly respond to common user questions, and provide context for its most popular data series[23]. The World Bank aims to develop further materials in this area including a toolkit that NSOs can share with potential users.

[21] http://www.opendataimpacts.net/engagement

[22] http://data.worldbank.org/about/open-government-data-toolkit/demand-for-od-engagement-tools#step_3

[23] http://blogs.worldbank.org/opendata

- **Application programming interfaces (APIs).** APIs enable machine-to-machine transfer of data from an Open Data platform directly to a user of that data in standard, machine-readable formats such as XML or JSON that promote interoperability. Development of APIs by or for NSOs will increase the ease with which users can access NSO data, and the ways that the data can be put to use. Examples include the recently launched API from the U.K. Office for National Statistics (ONS)[24] and the previously mentioned API from the U.S. Census Bureau[25].

- **Workshops and Training.** NSOs could develop or commission training for a range of potential users including, for example, journalists, researchers and students. This could cover both issues related specifically to Open Data, for example the use of APIs, but also wider areas such as training in statistical literacy. Workshops could also be developed around specific data products and resources. For instance, ISHN has been sponsoring outreach workshops to instruct academics and researchers about microdata products and techniques[26], but the same approach could be employed for other data programs.

- **Data publication calendars,** and adherence to publication release dates, as a trust-building measure.

- **Standards related to Open Data and open metadata**

- **Technical assistance.** Assistance to other government agencies that provide data to the national Open Data portal on data documentation, data quality review, anonymization techniques and similar subjects.

- **Regulations for other agencies.** On similar subjects as the point above, when the legal framework allows the NSO to enact and enforce such regulations.

This working paper makes the following recommendations in this area:

- Develop new methods of identifying the non-traditional users that Open Data will produce.
- Support new methods of encouraging the innovative use of data, for example through competitions.
- Commission and deliver a range of support to users including tools and guidance, APIs, training and metadata.

[24] https://www.ons.gov.uk/ons/apiservice/web/apiservice/home
[25] http://www.census.gov/newsroom/releases/archives/miscellaneous/cb13-tps37.html
[26] Sri Lanka: http://www.statistics.gov.lk/NewsInBrief/MOW_SL_Report.pdf. Rwanda: http://paris21.org/node/1597. Cambodia: http://adp.ihsn.org/node/1824.

7 Further Study

The working group and individuals who reviewed earlier versions of this working paper have identified the following related topics that merit further study:

- **Tools for data analysis and dissemination,** including data cubes, microdata, ad-hoc visualization and analysis tools, and georeferencing. Currently, the World Bank is conducting an assessment of data dissemination platforms that NSOs might use under Open Data programs. A report is due in June, 2014.
- **Metadata,** particularly, implementation of and experience with machine-readable metadata protocols such as SDMX and DDI.
- **Standards** for both data dissemination and APIs with the goal of increasing data interoperability, including a better understanding of the role of Linked Data and RDF.
- **Shared Ontologies** used to define and describe core statistical products and data series, building upon the development of standards by the international statistical community.
- **Application Programming Interfaces (APIs)** to provide standardized collection and transfer of census and survey data, building upon existing tools.

Appendices

1 NSO Study Working Group

Name	Organization
Agus Suherman	World Bank Consultant; formerly BPS-Statistics Indonesia
Oliver Fischer	Census Bureau, United States
Eduardo Gracida	INEGI, Mexico
Daniel La Buonora	Instituto Nacional de Estadistica, Uruguay
Federico Segui	Instituto Nacional de Estadistica, Uruguay
Vincenzo Patruno	Italian National Institute of Statistics
Carlo Vaccari	Italian National Institute of Statistics
Donath Nkundimana	National Institute of Statistics, Rwanda
Rajiv Ranjan	National Institute of Statistics, Rwanda
S. Altantseteg	National Statistics Office, Mongolia
L. Myagmarsuren	National Statistics Office, Mongolia
Laura Dewis	Office for National Statistics, United Kingdom
François Fonteneau	PARIS-21
Yoyo Chen	PARIS-21
Georgy Oksenoyt	ROSSTAT, Russia
Bill Joyce	Statistics Canada
Amparo Ballivian	World Bank
Thomas Danielewitz	World Bank
Olivier Dupriez	World Bank
Tim Herzog	World Bank
Barbro Hexeberg	World Bank
Iulian Pogor	World Bank
Matthew Welch	World Bank
Tim Harris	World Bank (consultant)

2 5 Stars of Open Data applied to NSOs

In 2010, the inventor of the World Wide Web, Tim Berners-Lee, proposed a 5-star model for linked Open Data.[27] This model was intended to encourage people, particularly government data owners, to adopt best practices for disseminating data without advocating prohibitive standards. While the original 5-star model is limited to technical elements (other organizations have proposed expanding the model for other aspects of Open Data), it is still useful for demonstrating how current dissemination approaches look from an Open Data perspective.

Each successive step in the 5-star model builds upon the previous one. The following table describes the 5-star model with examples of how NSO data might typically be disseminated at each level.

Stars	Description	Examples
★	Available online, openly licensed, in any electronic format	Statistics in non-machine-readable tables (GIF or JPEG), unstructured HTML, or embedded in PDF reports. Data is publicly available, but difficult to search, and must be re-entered by hand.
★★	Available online, openly licensed, in common electronic formats.	Data files in proprietary formats such as Excel, SPSS, SAS or STATA, which require special software and training.
★★★	Available online, openly licensed, in non-proprietary electronic formats.	Data files in open formats such as CSV, JSON, XML, DDI, SDMX, or structured ASCII
★★★★	All of the above, plus use of unique URIs (unique internet identifiers) to identify and define data.	Data files in Linked Data formats such as RDF, allowing data to be interlinked with other data files easily.
★★★★★	All of the above, and link to other data to provide context	Dataset pages provide machine-readable metadata, and link to standard definitions and related information

Note that each successive level builds on the previous one in ways that are not mutually exclusive. For example, NSOs can distribute data in non-proprietary formats such as CSV *and* proprietary formats such as STATA as many users will prefer the latter format.

[27] http://www.w3.org/DesignIssues/LinkedData.html

3 Summary of Select Microdata Access Policies

US Census Bureau

Microdata files from the Economic directorate of the Census Bureau, which manages surveys of businesses, are highly restricted, and not considered public use. The reasons for this are twofold. First, U. S. Law (Title 13) forbids the Census Bureau from publishing any information that can identify individuals or establishments. Second, the Census Bureau enjoys an unusually high level of trust among the business community, which it considers essential to the integrity of its work. Accordingly, access to microdata from enterprises is restricted to Census Bureau staff and to the Center for Economic Studies, which provides secure access to restricted-use microdata files to qualified researchers with approved products. Otherwise, almost no public use files are currently available (the exception being microdata from surveys of government agencies, which are made available upon demand).

In contrast, microdata files from the Demographic Directorate of the US Census Bureau, which manages several surveys of individuals and households, are routinely made available as Public Use Microdata Sample (PUMS) files. These data are heavily anonymized (discussed below) but otherwise contain unit-level responses and are made available under terms that are generally compatible with Open Data licensing (e.g., registration is not required for access, and no restrictions are placed on use). Surveys released as PUMS files include the American Community Survey (ACS) and the Decennial Census files.

PUMS files employ several anonymization techniques to ensure that identifying information is not disclosed. First, the files are actually random "samples" of the survey universe, such that one cannot ascertain with certainty if a particular individual or household is included in the sample (this would not be as effective with business surveys). Still, PUMS files still constitute substantial samples; the ACS survey contains over 3 million individual responses and almost 1.5 million household responses. This technique would be less effective in the economic realm, which consists of 7 million businesses compared to 317 million individuals.

Second, typical geographic identifiers (such as city and county) are replaced in PUMS files with "Public Use Microdata Area" (PUMA) identifiers, which represent geographic areas containing no fewer than 100,000 people each. Other anonymization techniques, such as top-coding and grouping outlying responses, are also employed.

The Census Bureau is assessing the possibility of providing PUMS files for a few of its economic (business) products as well, although these plans are very preliminary.

http://www.census.gov/main/www/pums.html

World Bank Microdata Library

The World Bank Microdata Library contains microdata from the World Bank and many other sources. The Microdata library is not considered to be Open Data. Anyone can access the library but may need to register or receive permission to access certain datasets. The Microdata library has five discrete policies that may be applied to microdata:

- **Direct Access.** Users are not required to register to access data. Terms of Use require proper citation, limit use to statistical and research purposes only, and forbid identification of respondents, redistribution without prior consent, and linkages to other datasets for the purpose of re-identification.
- **Public Use.** Users must first register to gain data access (there are no restrictions on registration). Terms of Use are the same as for Direct Access.
- **Licensed.** Users must request access and receive authorization before accessing each dataset, and must identify how the data will be used in the data request. Terms of Use are otherwise the same as for Direct Access.
- **Available from External Repository.** In this case, the microdata website functions as a portal to other repositories. Datasets are listed in the microdata website, but access is governed by the external repository with its own terms of use.
- **No Access.** Data access (beyond metadata) is not available.

http://microdata.worldbank.org/index.php/terms-of-use

Statistics Canada

Statistics Canada makes certain types of microdata available to researchers in a variety of ways including:

Public Use Microdata Files

These files are vetted for confidentiality and are made available without cost to users who sign a license agreement, which contains the following:

1. Statistics Canada hereby grants to the Licensee a non-exclusive, non-assignable and non-transferable license to use the Microdata file and related documentation for statistical and research purposes. The Microdata file shall not be used for any other purposes without the prior written consent of Statistics Canada.
2. Use of the Microdata file is limited to the Licensee. The Microdata file cannot be reproduced and transmitted to any person or organization outside of the Licensee's organization.
3. The Licensee shall not merge or link the records on the Microdata file with any other databases for the purpose of attempting to identify an individual person, business or organization.
4. The Licensee shall not present information from the Microdata file in such a manner that gives the appearance that the Licensee may have received, or had access to, information held by Statistics Canada about any identifiable person, business or organization.

5. The Licensee shall not disassemble, decompile or in any way attempt to reverse engineer any software provided as part of the Microdata file.

Research Data Centers

Secure on site access to certain microdata files is available to registered users whose projects are approved in advance. All output is vetted for confidentiality by Statistics Canada staff.

Real Time Remote Access

Eligible users who agree to strict terms and conditions are able to remotely submit programs which run against detailed analytic files. The output is automatically vetted for confidentiality. This service is provided on a cost recovery basis.

All access to microdata files is considered restricted in some sense and Open Data licenses do not apply. Standard and custom aggregate output, however, are considered open. These data files, when accessed via the Government of Canada Open Data Portal are governed by the open government license: http://data.gc.ca/eng/open-government-licence-canada.

When these files are accessed from Statistics Canada directly they are governed by the Statistics Canada open license: http://www.statcan.gc.ca/eng/reference/licence-eng.

IPUMS International

IPUMS-International offers free access to integrated census microdata and metadata for more than 70 countries to researchers and policy makers world-wide. The database, totaling over 500 million person records (2014), is updated annually with high precision household samples for the latest 2010 round censuses and for additional countries.

100+ National Statistical Offices have endorsed a uniform memorandum of understanding to facilitate access by means of a single license agreement managed by the University of Minnesota Population Center. Registration is required to assure compliance with the conditions of use license regarding statistical confidentiality, non-commercial usage, proper citation of publications, and remedies for violations. More than 8,000 researchers have registered for access, representing over a thousand institutions in some 120 countries.

http://www.ipums.org/international
https://international.ipums.org/international/international_partners.shtml

Italian Statistical Office – micro.STAT

The Italian Statistical Office makes certain microdata available for free to the public through its micro.STAT initiative. Currently, six datasets are available for download. As of this moment, user registration is required, and users must agree to cite datasets appropriately (following Creative Commons guidelines) and use the data only for statistical purposes. The public data files are anonymized sub-samples of the original data files intended for research projects.

http://www.istat.it/en/archive/public-use-micro.stat-files