Articles, special issue: ODSCSD

Open Government Data and Evidence-based Socio-economic Policy Research in India: An overview

Aurelie Larquemin
Institute for Financial Management and Research, India
Corresponding Author.
aurelie.larquemin@ifmr.ac.in

Jyoti Prasad Mukhopadhyay
Institute for Financial Management and Research, India
jyoti.mukhopadhyay@ifmr.ac.in

Sharon Buteau
Institute for Financial Management and Research, India
sharon.buteau@ifmr.ac.in

Public entities are one of the main producers of socio-economic data around the world. The Open Government Data (OGD) movement encourages these entities to make their data publicly available in order to improve transparency and accountability, which are the two important pillars of good governance. Thus, OGD by virtue of making quality data available can promote evidence-based public policy through empirical research. Therefore, in this paper we discuss the current status of OGD initiative in India, and feedback on the same from researchers working on India's socio-economic issues.

Larquemin, A., Buteau, S., Mukhopadhyay, J.P. (2016). Open Government Data and Evidence-based socio-economic policy research in India: an overview. *The Journal of Community Informatics*, 12(2), (Special issue on Open Data for Social change and Sustainable Development), 120-147.

Date submitted: 2015-08-07. Date accepted: 2016-05-09.

Introduction

In recent times, open government data (OGD)[1] has gained a lot of currency across the globe, thanks to the Open Government Partnership (OGP) initiated in 2011. More than 50 countries have committed to improving OGD access as a means for promoting important issues such as economic growth, transparency, development and governance. The launch of 'data.gov' in the US in 2009 and 'data.gov.uk' in the UK in 2010 mark the beginning of the OGD movement. Subsequently, other developing nations have joined OGP and are gradually catching up. As per the Open Government Working Group guidelines issued in 2007, OGD must be 'complete, primary, timely, accessible, machine processable, non-discriminatory, non-proprietary, and license free' (Open Government Working Group Guidelines cited in Yannoukakou & Araka, 2014, p.336).

Without doubt, OGD has important implications for research. Researchers in all areas rely on quality data to conduct their studies and test their research hypotheses. One of the first steps in any rigorous and systematic empirical research study is to determine what kind of data is needed to answer the research question being studied and whether the required data is currently available, and if not, how to collect it. If the chosen topic of research is fairly specific, then collection of primary data becomes inevitable for researchers. However, the cost of data collection, the time and other resources required to do so often pose serious challenges. Secondary data disseminated by various government departments under the OGD initiative appears to be the next best option available to the researchers, subject to their scope and quality.

OGD, with proper quality checks, can be envisaged as a viable avenue to provide data and to promote applied social science research in any country. However, in the context of developing countries like India, this option remains relatively scarce. Over the last decade subsequent to the enactment of Right to Information (RTI) Act in 2005, the Government of India has undertaken a number of initiatives such as the National Data Sharing and Accessibility Policy (NDSAP), an OGD portal in India (data.gov.in), the National Data Bank and DevInfo India. Little is known about the effectiveness of OGD for research in India. This study fills this particular knowledge gap in the Indian context. Against this backdrop, we examine in this paper awareness about OGD availability among researchers working on India-specific issues, use of OGD by sector and shortcomings of existing OGD.

In order to achieve our stated objective, we surveyed researchers and academics who had been conducting socio-economic studies in India using a structured online questionnaire. Additionally, we conducted interviews and a workshop with relevant stakeholders in India. Our results suggest low level of awareness among researchers about OGD. We find that

[1] To be considered 'open', data must be available as a whole, and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form. It must be provided under terms that permit reuse and redistribution including the intermixing with other datasets. The data must be machine-readable. Everyone must be able to use, reuse and redistribute — there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, are not allowed (Open Data Handbook, 2012). Open government data means data produced or commissioned by government or government controlled entities which is 'open' as per the Open Data Definition – that is, it can be freely used, reused and redistributed by anyone. Open data can also be generated by corporates by disclosing information pertaining to their business and related activities. Similarly, open datasets created by research institutions or individual researchers are termed as Open Research Data. In this paper we focus exclusively on open government data in socio-economic areas.

downloading large datasets, non-availability of metadata, and format of the datasets pose serious challenges to the use of OGD available in research.

The rest of the paper is organised as follows. In Section II we briefly discuss status of OGD initiatives in India. In Section III we describe how research can influence policy-making and, in particular, how OGD, through its effects on research, can inform evidence-based policy-making. Section IV presents the methodology we adopted to gather feedback and views on OGD from researchers conducting applied socio-economic research in India. Results are presented in Section V, and Section VI concludes.

Status of Open Government Data in India

Access to quality OGD is quintessential for effective policy-making in any developing country like India (Davies & Alonso, 2013) Open data is at the heart of open knowledge and transparency. Organizing Open Data Day on 22 Feb every year across various countries has added further momentum to global open data movement. The idea behind this particular event is to promote adoption of open data policies by governments at all levels: national, regional and local.[2] Embracing the importance of OGD for good governance, the Government of India (GoI) adopted a number of strategies which we discuss briefly next.

The first stepping-stone to OGD is to set up a sound statistical system. Towards this, the GoI constituted the National Statistical Commission in 2000. Subsequently the commission submitted a report in 2002 with recommendations for strengthening and improving India's statistical system.[3] This comprehensive report also identified the extent of critical data gaps in every ministry and government department. Based on the recommendations, Indian Statistics Act 2008 and Collection of Statistics Rules 2011 were enacted. Around the same time, in 2005, India enacted the Right of Information (RTI) Act which 'mandates timely response to citizen requests for government information'. Enactment of RTI was a commendable step towards greater transparency and good governance.[4] In fact RTI and OGD can act in tandem to achieve the same goal: 'to increase transparency of government by releasing information generated and collected by public funds in order for the citizenry to be benefit from its social and economic value' (Access Info, cited in Yannoukakou & Araka, 2014, p.336). In the same year, GoI constituted the National Knowledge Commission headed by Sam Pitroda which made several recommendations to improve India's knowledge network. One such recommendation was to enhance government data dissemination through a national web-based portal for certain key sectors such as agriculture, industry, water, energy and environment. Later, in 2006, GoI also introduced the National E-Governance Plan (NeGP) with an overall goal of making government services more efficient, transparent, reliable and accessible to a common man in India through Common Service Centre (CSC). In 2012, GoI went one step further in terms of active dissemination of government data by adopting the

2 See http://opendataday.org/

3 See Srinivasan (2003) for a critical appraisal of the report submitted by National Statistical Commission (2002)

4 Under this act a person can obtain required information by submitting a request to the Public Information Officer (PIO) of the respective government department.

National Data Sharing and Accessibility Policy (NDSAP). Following NDSAP, the OGD data portal http://data.gov.in[5] was launched later in the same year.

OGD is by and large released through public entities' portals, the dedicated OGD portal www.data.gov.in and also through an extended network of intermediaries such as research centres and other NGOs. At this juncture, it is worth mentioning that The Ministry of Statistics and Programme Implementation (MOSPI) adopted a pro-active role in data dissemination before the advent of RTI or NSDAP by publishing a national data dissemination policy in 1998. MOSPI has set up a metadata repository powered by the National Data Archive (NADA) software developed by the International Household Survey Network (IHSN). The metadata provided in the archive includes, survey methodology, sampling procedures, questionnaires, instructions, survey reports, classifications, code directories etc. While the metadata is available, to access the datasets, a fee must be paid to the Computer Centre, in charge of the dissemination of the collected data by the Ministry. Hence, although the metadata catalog is listed on the OGD portal, the database is not, and it is accessible only on payment. Therefore, such data cannot be classified under OGD. The process of payment is another obstacle to easy access of the data by researchers as payments are to be made by demand draft issued by an Indian bank.[6]

In 2004, Reserve Bank of India (RBI), India's apex regulator of banking services, made its internal database on Indian economy accessible to the general public. The RBI dataset, Database on Indian Economy (DBIE), is rich in terms of its content.[7] A few private initiatives which source information from government departments, organise and disseminate them under different thematic topics for free are also worth mentioning. Open Knowledge Foundation in India has launched a web portal, India City Open Data Census, to track openness of seven major cities in India in terms of annual budget, expenditure, election results, etc.[8] Another civil society initiative which is worth mentioning is IndiaGoverns which seeks to make development data accessible and useful for policy-makers, researchers and general public.[9] India Water Portal (IWP) is another such web-based portal dedicated exclusively to water management knowledge dissemination.[10]

---

5 The OGD portal gathered datasets from the following government entities as on April 2015: Comptroller and Auditor General of India(CAG) (16); Department of Atomic Energy (1); Department of Space (15); Lok Sabha Secretariat (100); Ministry of Agriculture (367); Ministry of Chemicals and Fertilizers (14); Ministry of Civil Aviation (3); Ministry of Commerce and Industry (17); Ministry of Communications and Information Technology (18); Ministry of Corporate Affairs (26); Ministry of Defence (23); Ministry of Development of North Eastern Region (1); Ministry of Drinking Water and Sanitation (MDWS) (16); Ministry of Earth Sciences (10); Ministry of Environment and Forests (10); Ministry of Finance (124); Ministry of Health and Family Welfare (199); Ministry of Home Affairs (244); Ministry of Human Resource Development (67); Ministry of Information and Broadcasting (13); Ministry of Micro, Small and Medium Enterprises (14); Ministry of Mines (26); Ministry of New and Renewable Energy (12); Ministry of Panchayati Raj (3); Ministry of Petroleum and Natural Gas (26); Ministry of Power (5); Ministry of Road Transport and Highways (133); Ministry of Rural Development (2); Ministry of Science and Technology (92); Ministry of Statistics and Programme Implementation (345); Ministry of Tourism (3); Ministry of Water Resources (557); Planning Commission (776);Rajya Sabha (154).

6 An online payment system to purchase MOSPI data is currently being implemented.

7 See http://dbie.rbi.org.in/DBIE/dbie.rbi?site=home

8 For more information see http://in-city.census.okfn.org/

9 See http://www.indiagoverns.org/. Currently it focuses only on Karnataka.

10 See www.indiawaterportal.org

The effects of research results on policy-making: How OGD could increase evidence-based policy-making.

Do research findings lead to evidence-based policies?

To assess how OGD can promote evidence-based policy-making, it is useful to understand how research findings can lead to evidence-based policies. The common assumption about research is that the findings should have a direct impact on the decisions made by policy-makers and practitioners. However, in reality, a deeper analysis is required to discern the various paths in which results from research penetrate into policy and practice. This issue is not new, however, and with the progress of technology to facilitate dissemination of research, as well as easier access to data through OGD, the topic merits further discussion. The emergence of evidence-based public policy making can be traced back to 1833 when a study was conducted to show that education did not reduce crime (Guerry, 1833). Much later in the 1970s, it became evident that research findings often failed to have an impact on policy-making. Subsequently research was done to identify the failures in this process. What researchers came to understand was that their findings were only a minor component in the equation leading to policy-makers' actions. Other elements contribute to the policy-making process: political interests, ideological convictions, concerns about resources (staff and budget) to implement new activities, bureaucracy, weight of tradition, etc. Research findings were rarely acknowledged and often drew little attention of policy-makers. Hence, the use of research appeared more complex than initially considered. This process was conceptualised and refined over the years. Whiteman in 1975 describes a two-dimensional perspective on research use in policy-making: a concrete and instrumental process where research findings are a fundamental component of public policy, and a conceptual and indirect path to influence policy-making, by giving policy-makers a deeper understanding of issues in their field, new ideas or motivation, and a new perspective on the targeted issues. Greenberg and Mandell (1991) in their work on research utilisation in policy-making adapted and refined this framework by considering that in both cases research could affect policy-makers in different ways starting from a substantive manner to a more influential way or in a strategic purpose. In this model, each dimension– concrete and conceptual – is envisaged as a continuum.

This same framework was also adopted by Nutley, Walter and Davies (2007) in their study on how research can inform public services. The authors also argue that evidence-informed or even evidence-aware policy would be a better description of the aspirations for the role of research in the policy making process.

Several efforts have been put in to identify this complex process and to eventually improve the use of research in public policy making. Webber (1991) in a study titled 'The distribution and use of policy knowledge in the policy process' argue that policy knowledge is not effective if it is not shared. It should be efficiently and extensively communicated and explained to policy-makers in order to influence policy decisions. To achieve this, a better understanding of the barriers to research affecting policy, as well as the research and policy connection are essential (Watt, 1994).

Table 1: Research use as a two-dimensional continuum

	Substantive	Elaborative	Strategic
Concrete	Research shapes the core of a decision or an issue	Peripheral use of research to further refine a position	Research is used to justify a position that has already been adopted
Conceptual	Research shapes a core orientation towards an issue or a basic understanding of the issue	Peripheral use of research to further refine an orientation or understanding	Research is used to confirm an orientation or an understanding that has already been adopted

Source: Adapted from Greenberg & Mandell (1991); Nutley, Walter & Davies (2007).

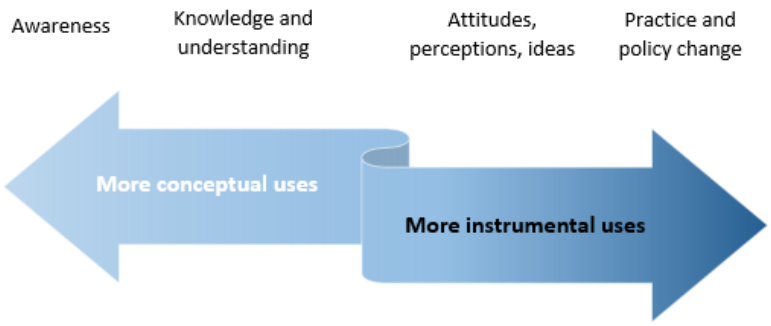

Figure 1: A continuum of research use (Adapted from Nutley, Walter & Davies (2007))

Source: Adapted from Nutley, Walter & Davies (2007)

- Barriers to research affecting policy

It is now becoming increasingly clear to the research community that the context in which the research is being conducted has an influence on the uptake of the findings by policy-makers (Nutley et al., 2007). Some conditions seem to predispose policy-makers to take research findings seriously. Indeed policy-making in developing countries have specific characteristics and the way to conduct and consider research can affect the uptake and adoption of research findings by policy-makers (Carden, 2009). According to Carden (2009), elements like the number of points of access through which research findings can flow, the openness of the political system to the entry of new ideas and the democratic nature of decision-making determine the degree to which research can influence policy making. Carden (2009) study the consequences of 23 research projects funded by the IDRC and found evidence that development research, if done correctly, can improve public policy and help accelerate development progress. These findings highlight the fact that well-designed research, if properly executed and disseminated, can influence public policy. However, on many occasions development research frequently fails to register any apparent influence in

developing countries. Carden (2009) also identify certain stylised facts about research and policy-making environment in developing countries that could explain this failure. First, often policy-makers have less autonomy. Second, staff turnover in research organisations and in government is high which weakens the link between research and policy-making. Third, developing countries often lack the intermediary institutions that may influence research to policy. Lastly, implementation challenges are greater, both for research activities and for policy- making. Carden (2009) conclude that researchers in developing countries often lack access to required data and hence they might have to resort to the creation of a database through primary surveys. However, even if all those issues were addressed, another aspect to consider is scientific uncertainty. This can lead to distortion and lack of clarity in policy-making. If policy- makers receive disaggregated or opposite information from many different research projects, they may be unable to assess what action to take.

Several other studies reach similar conclusions. For instance, Court and Young (2003) undertake a comparative analysis of 50 case studies collected during the first phase of the Global Development Network (ODI, 2003). The authors identify gaps in the theory of the path of research to influence public policy due to a failure to take into account specific characteristics of developing countries. According to them, 'the key issue affecting uptake was whether research provided a solution to a problem. Policy influence was also affected by research relevance (in terms of topic and, as important, operational usefulness) and credibility (in terms of research approach and method of communication)'. Court and Young (2003) also highlight the importance of a clear and well-conceived communication strategy and strong advocacy efforts from the very beginning, relating to the local context and concepts familiar to local policy-makers.

Therefore, international development agencies and other research funders are placing increasing emphasis on the need to communicate research evidence to policy-makers, taking into account not only the demand side of evidence from research, but also supply side of making evidence accessible to policy-makers in a comprehensive and timely manner. It is recommended that researchers ought to pay greater attention to their communication and dissemination strategy and ensure that it is made available to policymakers when the research would be the most useful.

- Assessing the research–policy connection

Following the International Conference on Evidence-Informed Policy-Making held from 27 to 29 February 2012, in Ile-Ife Nigeria, Newman et al. (2013) prepare a paper titled 'What is the evidence on evidence−informed policy-making? Lessons from the International Conference on Evidence−Informed Policy Making'. In this paper they present an updated model which explains how research results can affect policy-making.

Lavis et al. (2010) study the engagement of researchers in bridging the gap between research findings and policy-making. The authors identify three sets of activities: providing systematic reviews of the research literature to their target audience, giving access to a searchable database of research products on their topic, and establishing or maintaining long-term partnerships related to their topic with representatives of the target audience. They survey 308 researchers who has been conducting research on one of four health issues critical for the achievement of Millennium Development Goals (prevention of malaria, care of women

seeking contraception, care of children with diarrhoea and care of patients with tuberculosis) in each of ten low- and middle-income countries (China, Ghana, India, Iran, Kazakhstan, Laos, Mexico, Pakistan, Senegal and Tanzania). Their results show that important research findings often remain poorly disseminated: less than half of the researchers surveyed reported that they engaged in one or more of the three potential dissemination activities: 27% provided systematic reviews of the research literature to their target audiences, 40% provided access to a searchable database of research products on their topic, and 43% established or maintained long-term partnerships related to their topic with representatives of the target audience. Among the factors explaining the respondents' engagement in these activities were (i) the existence of structures and processes to link researchers and their target audiences; (ii) the stability in their contacts; (iii) having managers and public (government) policy-makers among their target audiences.

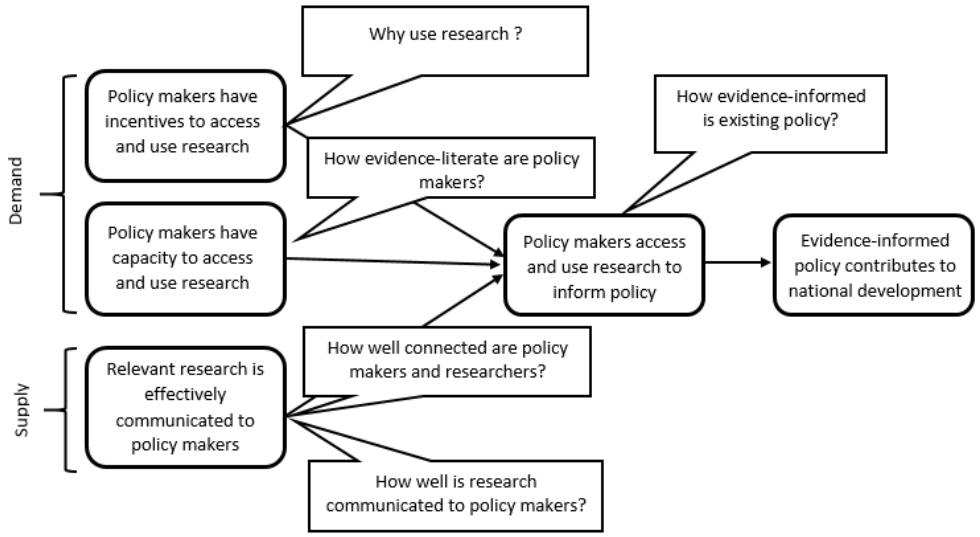

Figure 2: Factors contributing to evidence-informed policy (Newman, Capillo et al. (2013))

Can OGD promote evidence-based policy making?

There are various hypotheses on how open data in general and OGD in particular may affect policy-making. These hypotheses are considered for OGD in particular. The ODDC Conceptual Framework (Open Data in Developing Countries, 2013) has outlined three hypotheses, each of which have a distinct focus area:

Table 2: How open data (OD) can influence policy (ODDC Report, Researching the emerging impacts of Open Data: ODDC Conceptual Framework, July 2013).

Hypothesis	**Key focus**	**Key disciplinary traditions/ Streams**
OD can bring about greater transparency in government, which in turn brings about greater accountability of key actors to make decisions and apply rules in the public interest.	The State (political domain)	Political science, public administration, legal studies

OD can enable non-state innovators to improve public services or build innovative products and services with social and economic value; open data will shift certain decision making from the state into the market, making it more efficient.	The Market (economic domain)	Economics, business models, regulation
OD can remove power imbalances that resulted from asymmetric information, and will bring new stakeholders into policy debates, giving marginalised groups a greater say in the creation and application of rules and policy;	The 'Excluded' (social domain)	Social science, community informatics

- The State

Many studies on OGD tend to argue that the release of OGD will improve transparency and accountability of public authorities, and will eventually lead to better governance. (Arzberger et al., 2004; Chun et al., 2010) The publication of data increases transparency in governance and its accountability, which generate confidence in government's actions. This encourages public engagement which in turn leads to more efficient policies. According to Chun et al. (2010), the OGD movement is an integral part of governance, and the release of such data tend to act as a bridge between citizens and public authorities. Therefore, OGD can result in better governance and increased accountability of public authorities. In May 2015, at the University College London's Department of Information Studies (DIS) Research Symposium: Open Data and Information, Shepherd (2015) in her address emphasises the fact that simply publishing data would not be enough to achieve the stated objectives. Additional endeavour is needed to make raw data usable and reusable with the support of an organisation which will ensure non-redundancy of efforts and consistency and will guarantee data sharing, data integrity and quality. Countries are therefore required to build this much-needed system to ensure that OGD can have the expected impact, including on evidence-based policies.

Public bodies are among the largest collectors and producers of data in many different domains. These data domains range from traffic, weather, geographical, tourist information, demographic statistics, taxes and business, public sector budgeting and performance levels, to all kinds of data about policies and inspection (food, safety, education quality, etc.). The OGD movement in most cases assumes that public agencies are ready for an open governance process. One of the inevitable fall outs of such a process are discussions, discourses, debates and exchange of ideas, views and inputs. However some studies have shown that in several countries the political class can be reluctant to adopt open government measures. Janssen et al. (2012) list the reasons for this reluctance which includes the shift from a closed to an open system of governance that can have significant impact on relationships between the Government and the civil society. Some remain sceptical about the expected positive outcome of the OGD initiatives compared to a system with barriers, and so far no systematic research is available that addresses this concern.

- The Market

Shepherd (2015) points out that in countries where the OGD movement has been initiated early on and is continuing successfully, it has received strong support from the political class.

She cites the example of OGD in the UK pointing out that politicians argue that OGD 'would benefit the UK economy by creating jobs and stimulating innovation and at the same time increase transparency and accountability […], empower the citizens' public participation' and improve public services. However, the return on investment from open data in general remains unclear, since 'open data has no value in itself; it only becomes valuable when used' (Shepherd, 2015).

A multiplication of studies regarding the potential economic gain following the implementation of Open Data policies can be observed. McKinsey Global Institute in a report published in 2013 suggest that open data could have an important economic impact on societies. They claim an estimated USD three trillion in annual economic terms that could be unlocked across seven domains. These benefits include increased efficiency, development of new products and services, and consumer surplus (cost savings, convenience, better-quality products), without quantifying social benefits. They estimate the economic impact of improved education (higher wages), but not the benefits that society derives from well-educated citizens. They estimate that the potential value would be shared by the United States (USD 1.1 trillion), Europe (USD 900 billion) and the rest of the world (USD 1.7 trillion). The study clearly brought out the fact that, to reach a potential economic or political benefits, Open Data initiatives need to focus not only on release of data but also on the infrastructure, developers and researchers to make the data openness a reality and useful for addressing some of the development challenges faced by a country. This conclusion is valid for OGD initiatives, as part of the broader open data initiatives. Third parties, including researchers, have to be involved and they should play an essential role in cataloguing, cleaning, and other data related activities. They have to perform analyses to uncover valuable insights. Releasing metadata will make open data, including OGD, more usable and comprehensive. Investments in key skills including the ability to perform analyses, creating useful reports and tools, and incorporating data and results into managerial decision-making processes are the other necessities to make this movement successful.

- The 'Excluded'

There is consensus that OGD needs researchers to maximise its potential in terms of better governance, accountability, economic benefits, etc. It is important to point out that researchers also benefit from OGD initiatives. First, OGD offers opportunity to conduct new research studies by giving access to new datasets. Using OGD the researchers are also afforded the opportunity to identify new problems and conceive new research studies. According to Guo Xu (2012), among the benefits of open data initiatives on research is that the publication of data helps avoid redundancy and therefore waste of resources. The publication of data following best practices and open data standards avoids researchers repeating the same procedures (cleaning datasets, compiling, merging, formatting, etc.). There is no reason why this should not apply to OGD as well –open data and OGD allows for a better allocation of research resources, which are often scarce. A database collected for one study can feed another researcher's work on completely different problematics. Secondary data can be the key for important research work and OGD initiatives aim to make more secondary data available. .Groves (2012) in his paper on open science and reproducible research argues that data sharing can greatly increase dissemination, meta-analysis, and understanding of research results; it can also aid confirmation or refutation of research through replication, allow better implementation of research findings, and increase

transparency about the quality and integrity of research. Hence open data in general and OGD in particular, by increasing transparency and replicability of the research studies, should reinforce scientific rigour.

OGD and openness in research could stimulate research activities and outcomes, but this virtuous circle depends also on the willingness of researchers to share openly the data they use and their findings. While examining this issue, Azberger et al. (2004) opine that the openness of data is a way to ensure that both researchers and the public receive optimum returns on the public investments in research. It is also a means to build on the value chain of investments in research, and its data resource with the underlying principle being that 'publicly funded research data should be openly available to the maximum extent possible'. Although many calls for proposals and study datasets have been published to mark the beginning of this movement, many have concerns regarding the open publication of their data and results. Shepherd (2015) rightly points out that researchers may be reluctant to release data, as there is a fear that original ideas and research will be stolen or misunderstood. Other studies also insisted on the fact that the idea of open access to data in research is professed by many but not practised by many (Hamermesh, 2007). The advantages for the research community and beyond generally fail to outweigh researchers' fears and 'costs' to create the supply of data and replicable results (Anderson et al. 2008).

Versbach et al. (2013) use a dataset of 488 economists from the top 100 economics departments and the top 50 business schools, and provided evidence of the status quo of data sharing or data access facilitation in economics. Using an ordered probit regression they identified that the likelihood of sharing dataset is positively associated with a number of factors such as sharing of other material, being full professor, and being affiliated with a higher-ranked institutions. In another study Piworar (2011) examine 11603 articles published between 2000 and 2009 that described the creation of gene expression microarray data. Using a multivariate regression, Piworar (2011) find that authors were most likely to share data if they had prior experience of sharing or reusing data, if their study was published in an open access journal or in a journal with a relatively strong data sharing policy, or if the study was funded by a large number of (full form) NIH grants. Also the use of OGD and openness in research could greatly benefit young researchers and PhD students in the learning process.

Few studies have been done to assess the obstacles faced by researchers, both in India and abroad in accessing socio-economic databases collected by the government. It is important to assess whether the cost and payment methods are the only barriers encountered by the researchers.

Chattapadhyay (2014) conducted a study to assess research and advocacy organisations' difficulties in accessing OGD pertaining to India. One of the key challenges identified is the fact that most of the data collected by public authorities are not made available anywhere in digital format. The flows of information emanating from the prevailing reporting system between local, state and central public authorities prevent the publication of many datasets in a timely manner. Bureaucratic hurdles and reluctance to innovate also prevent the publication of original databases. There is a need, to overcome these fault lines, to invest in the capacity of the Government agencies and reinforce their motivation to publish original data, and to develop direct interactions between data producers and data users. This project also highlights the danger generated by the lack of publication of government data, to create a space for a data reselling industry and a closed community of re-users of data. Public

authorities are rarely aware of the existence of such 'data black market'. Data intermediary organisations consulted for the study mentioned downloading OGD almost always from the government websites and almost never used the OGD portal, mostly because they were used to collect these datasets from the public authorities' websites long before the OGD platform was launched and they continued doing so. They also highlight that some OGD come most often in closed formats (that is, PDF files and HTML tables), even if recently efforts have been initiated to make datasets more researcher friendly. This poses an important barrier in converting data into formats amenable to rigorous analysis.

Wright et al. (2010) identify the benefits and challenges of OGD in India. First benefit identified is for the GoI itself: the data becomes more accessible for internal use and government's own information gathering and processing procedures can be improved as and when incorrect and outdated data are identified. Second, this may be one step forward towards fulfilment of its commitments towards the Right to Information (RTI) Act. Among the challenges identified, upgrading the entire infrastructure of information gathering, processing and sharing (which is currently being implemented, e.g. New DGCIS portal) tops the list. Insufficient standardisation is being discarded, as well as the need for system and semantic interoperability. Currently there is no use of common standardized formats and software standards, and different departments are gathering different information under the same heading, or the same information under different headings which make data consolidation extremely difficult.

The literature available allows the identificaton of a theoretical linkage among OGD, research findings and evidence-based policies. At the beginning of this virtuous process are the researchers and their ability and willingness to use OGD, a mandatory step for the next ones to take place. However there is a dearth of literature focusing on the demand-side of OGD.

Methodology

Our study seeks to understand the awareness and use of OGD by researchers in India by examining responses to the following questions: (1) Do researchers know about OGD? (2)Do they use OGD? If so how, if not why? (3) What are the types and sources of open data used and why? (4) What is the perceived quality of OGD? (5) What are the key concerns about using OGD for research? (6) Are there any gaps in available OGD?

Research design

Due to the nature of our institution, activities and network, we delimited our study to socio-economic research in India. We collected data and triangulated results from the following sources:

(1) A workshop was conducted with stakeholders from the GoI and researchers (Indian and international) working with Indian datasets.

(2) An online survey. The survey was conducted among a sample of researchers selected purposively from our extensive network of researchers and academics.

(3) Interviews.

From previous experience, we were aware of the difficulty in gathering information from government representatives through an online survey. We therefore started this study by

inviting government officials to a workshop with researchers from various universities around the world who have been working on Indian socio-economic issues empirically. To keep the discussions focused, we selected specific topics which included status of data availability on industry, micro, small and medium enterprises (MSME) and employment in India.

A white paper had been prepared beforehand by Asher and Novosad (2015) to gather the views of the academic community on industry data in India and the results so obtained were used to set the agenda for the workshop. The paper was shared with official statisticians and other participants a few days before the meeting for their comments and feedback on the same. A presentation on the results was made by Dr Asher to initiate discussions. The workshop was conducted to identify issues with industry data availability and quality in India, and scope for making it a part of OGD available in India, Hence, OGD and closed industry data were discussed during the workshop. For example most industry data in India is accessible upon payment and therefore cannot be considered as OGD. However, scope for making such data freely accessible was one of the points discussed during the workshop. Therefore, data not yet available as OGD were also included for researchers to explain why the payments modalities and other related issues can pose obstacles to applied researchers.

Participants at the workshop were representatives from the National Statistical Commission, from the Directorate General of Commercial Intelligence and Statistics, representative of the chief Statistician of India, from the Economic Statistics Division at MOSPI, representatives of the Sampling and Official Statistics Unit (SOSU), ISI Kolkata, formerly members of RBI, CSO, NSSO, Representative from World Bank Delhi, ADB India and ADB Manilla, a representative from the Indian Council for Research on International Economic Relations (ICRIER) and from Economic and Political Weekly (EPW), IGC India Central representatives and IFMR representatives.

Professors with a track record of research on Indian socio-economic sectors were identified and those working specifically on industry, MSME and employment issues in India were invited to take part in the discussions. Professors joined the workshop from Oxford University, University of California San Diego, Pennsylvania State University, from the University of Michigan, University of Delhi, Harvard University, and the City University of New York.

The discussions were divided in sessions, with each session introduced by a short presentation from a participant on his/her work regarding Indian datasets and then opened to discussions. Each session was facilitated by a participant familiar with the specific topic discussed. Three professors and three official statisticians made presentations.

The topics discussed in depth were:

- What India should do to improve the status of data availability on industry in the country;
- The sectors in the Indian economy which would benefit for having more data available and more research studies being conducted;

- How critical it is for researchers in these sectors to have access to micro-level data, and the need for the Indian authorities to allow researchers accessing this confidential micro-data;
- The latest efforts and initiatives from the public authorities regarding data collection and data dissemination;
- The difficulties to generate panel data considering the actual design of the data collection process;
- Confidentiality issues related to dissemination of OGD;
- The limitations in dissemination due to the regulatory framework; and
- The organisation of data collection and the multiple players involved on a similar topic.

All presentations were shared with the participants and organizers in PowerPoint® and notes were taken by four reporters. The entire workshop discussions were also audio-recorded.

To complete and confirm the insights from the workshop, we designed a detailed 61-question online survey questionnaire, using the online software Survey Monkey®. We focused on researchers in India and abroad working on socio-economic issues in the Indian context. We reached out to 28 among the top universities and schools in India actively engaged in research on socio-economic issues.

We contacted faculty members and PhD candidates at the national level through personal emails and also contacted their administrative division to ensure internal spread of the survey link. In total 411 emails were sent, including 274 to Indian nationals. We also shared this survey with 137 researchers and PhD students located abroad who had been conducting studies on socio-economic issues in India, based in 71 top ranking universities and schools.

Our fellow researchers at international research institutions such as the World Bank, Global Development Network (GDN) and the International Monetary Fund were also contacted.

We received 70 fully completed responses after sending reminders periodically.

The first questionnaire appeared to be too long and too detailed; some questions were confusing and ambiguous due to a lower level of knowledge about OGD among the respondents than we had expected. We chose polar questions and bipolar scale responses to facilitate the data analysis given the time allocated to the study. However, unstructured response formats were also used to collect the respondents' opinion. Based on the answers obtained to these questions, more guided multiple-options might have been more appropriate for several items asked. Our initial results led us to restructure the questionnaire, making it shorter and sharper, with multiple-options questions based on the preliminary results. A second survey was then conducted among the initially contacted researchers who had not responded to the first survey and nine answers were collected.

Table 3: Summary of responses received

First survey: Emails sent	**411**	
	National: 274	International: 137
Responses received	**70**	
Second survey: Emails sent	**341**	
Responses received	**9**	

We completed our data collection with two interviews through email with two official statisticians who could not take part in the workshop. The online interviews focused on the following aspects:

- Why is your institution making data more available/open to the general public? Is it an obligation regarding the legislation (Right to Information (RTI) Act, the National Data Sharing and Accessibility Policy (NDSAP)) or/and also a policy of your institution?
- Who do you think/know are looking at and using the data your institution publishes?
- Does your institution benefit from publishing its data? How? (Feedback from users, external research projects, etc.)
- What are the next steps for your institution regarding data publication? (For example, making the data available online, making data available for free, publishing more data, in other formats, on other websites, etc.).

The workshop helped us immensely to understand the challenges in using OGD for research and identifying potential solutions during the discussions between academics and official statisticians. The survey responses gave us insights on the level of understanding and use of OGD in India. They also confirmed some challenges that had already been identified in the workshop.

Limitations

In this study we focus on OGD released at the national-level while bearing in mind that various departments of state governments and local governments also release public data which are outside the purview of this paper.

What about other open data sources? This study focuses on the OGD released at the national level and by public entities. There is a wider OGD network available in India with many public institutions at sub-national levels who publish data. There are NGOs, think-tanks and other data intermediary organisations who also publish data gleaned from public entities. Such types of data are outside the purview of this paper. However, studying these data sources can certainly provide valuable lessons and insights on OGD in Indian context.

The methodology adopted in this paper has some limitations. First, our methodology fails to control ‘nonresponse bias’ emanating from heterogeneity between respondent and non-respondents. Second, our study sample potentially suffers from sample selection bias due to convenience sampling since the study was conducted among researchers in our extended

network. Third, the workshop we mentioned earlier did not focus solely on government data openly accessible and their usage in applied research, though more emphasis could have given on this particular issue. However, the purpose of the workshop was also to generate an interaction and dialogue between researchers using Indian OGD and the government officials in charge of data collection in an attempt to push towards more data being made available in the public domain.

Findings of our exploratory study

Our online survey received responses from faculties, PhD scholars or post-doctoral and research fellows which amounted to 79 completed surveys. Our respondents were actively involved in research studies with 75% of them having been involved in a socio-economic research project pertaining to India over the last two years, in areas as diverse as anthropology, finance, development, environment, etc. 78% of the researchers were involved in the design of the study, and 83% of them had used secondary data for their research.

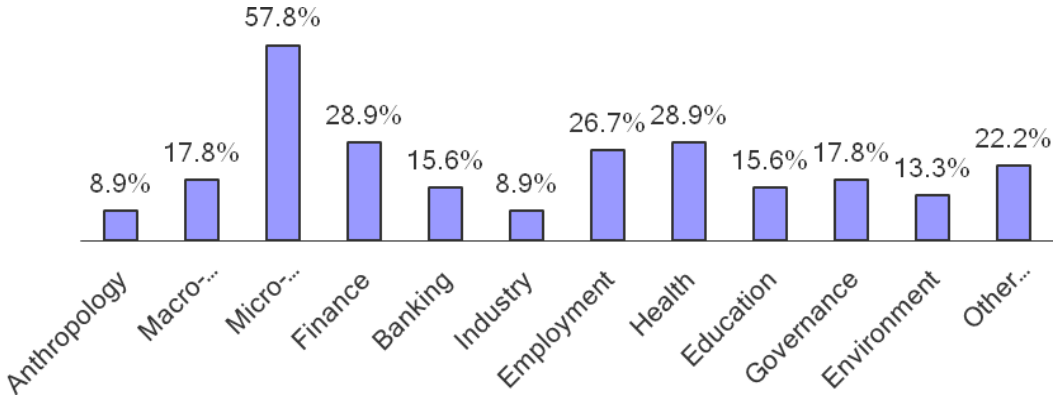

Figure 3: Have you been involved in a socio-economic project over the past two years? If Yes: What was the main domain of research? (n=79) (Online survey results)

The relationship between researchers and OGD in India: do they know and use OGD?

At the initial stage of our research we were under the impression that researchers were well aware of the OGD movement and such data availability in India, on either ministries' portals or ad hoc web based data platforms. But from our survey responses we realized that OGD is still not very popular among the research community. Only 57% of our respondents were aware of what OGD is.

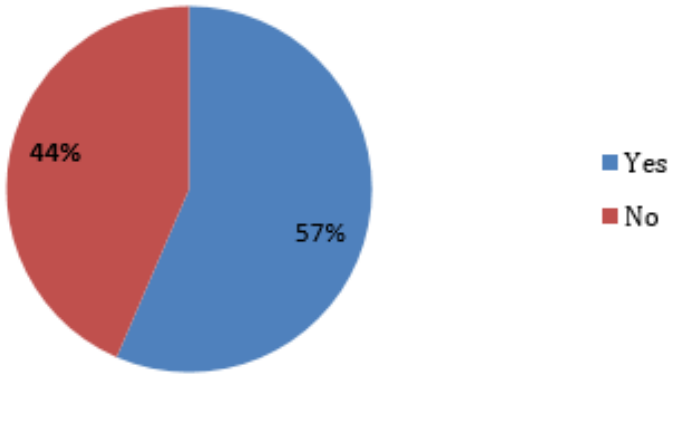

Figure 4: Have you heard of open government data? (n=79) (Online survey results)

Although the degree of knowledge among researchers familiar with OGD in India is relatively low, among those who had responded positively to this question, 25% and 38% considered that they had a low and average understanding of OGD respectively. Only 32% assessed their knowledge as satisfactory and only 6% felt they had extensive knowledge of OGD.

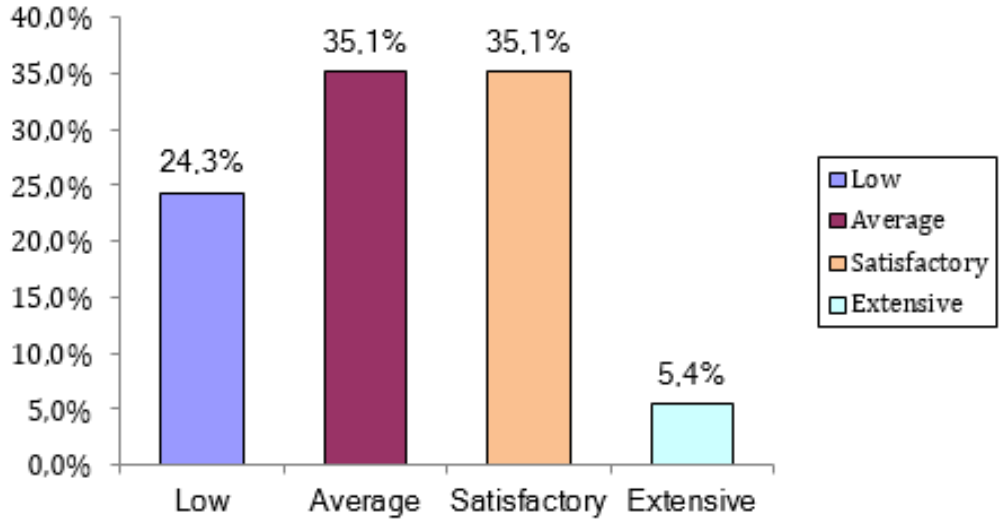

Figure 5: How would you define your knowledge of open government data sources? (Online survey results).

Among the researchers familiar with OGD, only half of them had consulted the official OGD portal, while 88% of them have already consulted data from the various ministries' portals. It would have been interesting to establish the reasons for this apparent low level of OGD use among researchers and to find out if it is due to a lack of awareness or for reasons intrinsic to the website.

The respondents familiar with the official OGD portal appreciated mostly the easy access and the fact that such data was not available elsewhere to the best of their knowledge. Considering that most, if not all, the data published on the official OGD portal is also available on their respective public data collectors' websites, a low level of OGD publication awareness in India can be suspected.

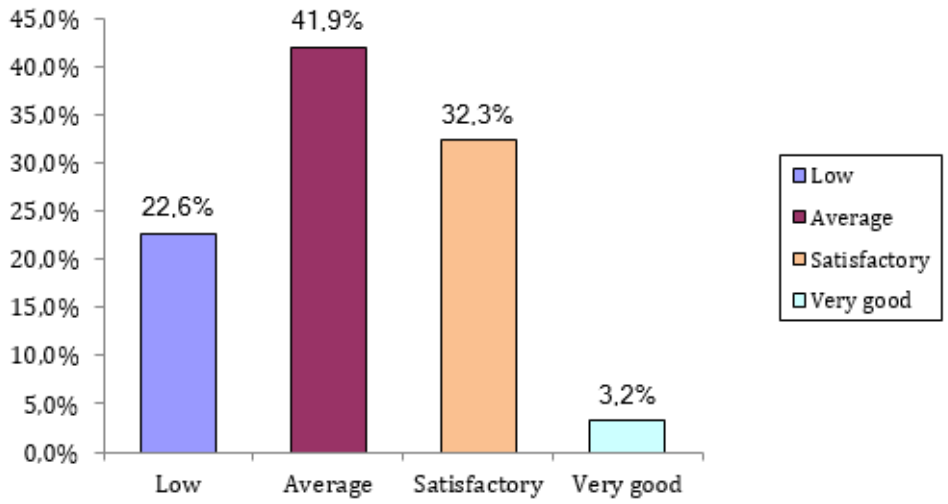

Figure 6: How would you rate the quality of open government data available in the Indian context? (n=79) (Online survey results)

Most of the respondents declared that they judge the quality of OGD available in India to be average (42.0%) or satisfactory (32.3%).

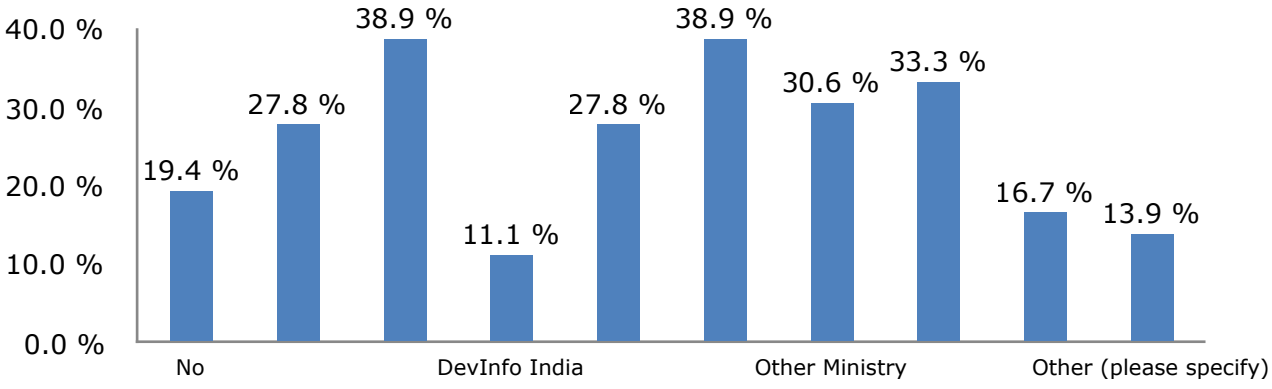

Figure 7: Have you ever worked with secondary OGD for socio-economic research purposes? (n=79) (Online survey results)

20% of the respondents have never worked with OGD in a socio-economic research project. Interestingly we found that while 39% of the researchers had used OGD in such circumstances from the website of the Ministry of Statistics and Programme Implementation (MOSPI), only 28% have done so from the official OGD portal. This might be due to the fact that some datasets are referenced or indexed on the OGD portal through only a link to reach the dataset, and hence sending the user ultimately to MOSPI website to access the actual data.

The share of researchers using secondary data collected by public authorities is higher than the numbers presented above. Of the respondents who had been involved in a socio-economic study in India in recent times, 80% used secondary data and they declared that they gathered it from public sources, but not only from open sources. A majority (60%) of them had to pay a fee to obtain the data (at a cost exceeding USD 100 for 70% of them). This shows tremendous potential for more public data to be released on the OGD portal and to be used by researchers. These datasets, despite being paid for, were not accompanied by any metadata as reported by 70% of the respondents. No additional services were offered by 'pay-to-access' portals compared to the OGD portal (metadata, troubleshooting/Q/A website). So there is no reason as to why researchers would not choose no-cost accessible datasets over data they have to pay for.

Advantages and challenges of using OGD in research

In our survey, we dedicated a section and a list of questions to understand researchers' reasons for using or not using OGD in their research to assess the challenges they had been facing and what their suggestions would be to improve OGD to maximise their use for socio-economic research. The same approach was adopted during the workshop and researchers and official statisticians discussed the difficulties they faced, and the suggestions they would make while public authorities presented their methodologies and the latest changes and improvements put in place in this matter.

- Findings from the on-line survey

Among the main advantages identified by the survey's respondents is the easy access of OGD if they are available. The platform data.gov.in allows one to quickly identify the database one is looking for. Many ministries and public institutions are also investing in making their platform more user-friendly and the datasets more accessible. Another advantage frequently pointed out is the fact that the access to these datasets is free. As previously mentioned, while exploring the OGD platform and other ministries websites, it was noticed that many key data on socio-economic research is only listed on these open portals but remain accessible only after paying a fee. Many datasets (e.g. MOSPI collected datasets) are by and large accessible on the payment of requisite fees.

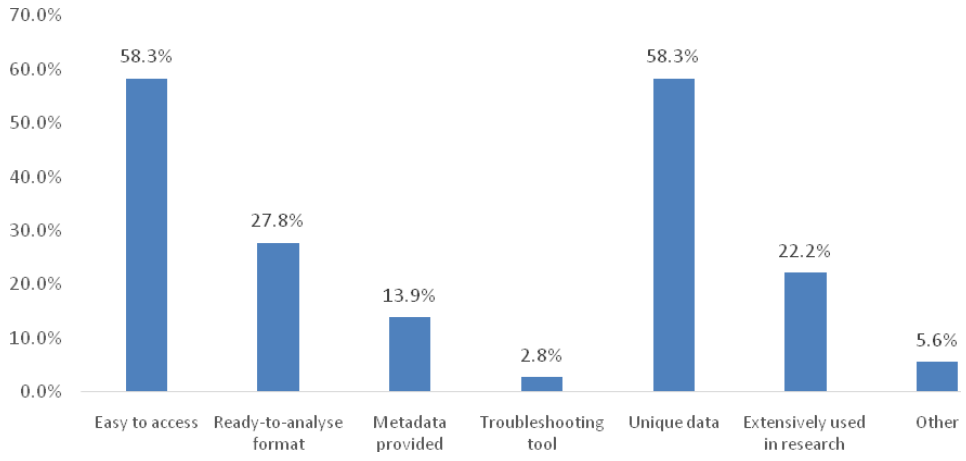

Figure 8: What was the most positive aspect you found about the portals consulted and the data displayed? (n=79) (Online survey results)

The formats of OGD publication and dissemination have been unanimously criticised. Current formats are not consistent across datasets and sometimes across years, and require a fair amount of researchers' time to make them amenable to analysis, for instance when they are delivered only in paper or in PDF format.

From the data, there seems to be no concern about the intrinsic quality of OGD. But many researchers expressed concern over the slow pace at which the public authorities make certain datasets available. Consequently, many datasets soon become outdated. For instance data from the Census 2011 have been published only recently, and hence many studies conducted over the past couple of years had to rely on data from the Census of 2001.

Many faced issues with downloading large quantities of data at once, and regretted that the source and data collection methodologies were often unclear and are far from comprehensive. Moreover, it was hard to get any support from the official side, while some lamented the lack of standardisation among datasets. A minority of the researchers who had consulted these OGD sources did not work with those datasets eventually (18%).

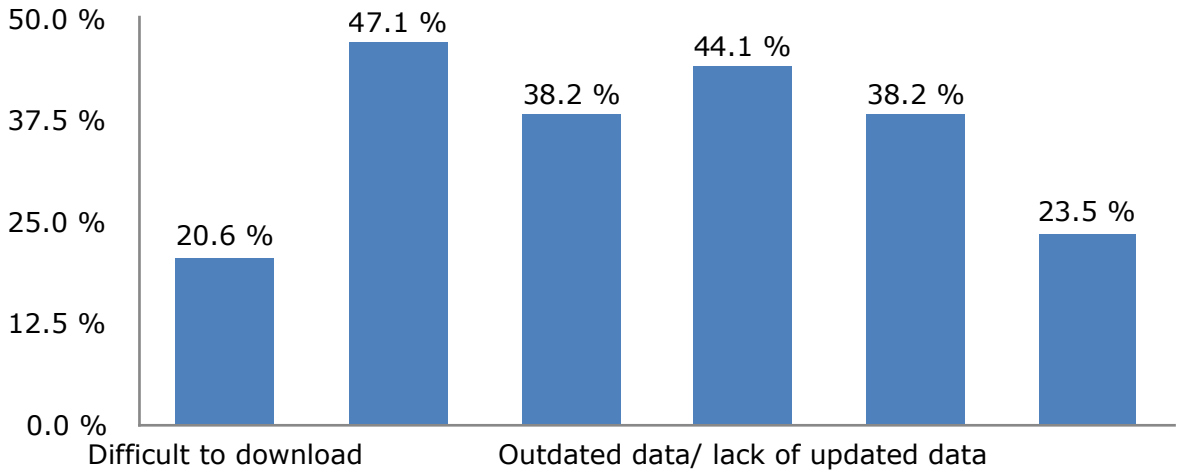

Figure 9: What was the most negative aspect you found about the portals consulted and the data displayed? (n=79) (Online survey results)

Despite their current level of usage, an improvement could be expected and researchers might be willing to work with OGD datasets in the future if some necessary improvements were carried out in dissemination practices. These necessary changes include the provision of metadata and the possibility to reach out effectively for troubleshooting. The format of dissemination has to be standardised and be made more user-friendly. 65% of the respondents familiar with OGD judged their quality as average or below par.

Another challenge that brings a consensus among the data users surveyed is the absence of good quality or even total absence of metadata associated with OGD. The lack of information regarding the public entities' methodology for their data collection is an important issue. Understanding the available data is of utmost importance for researchers to be able to use data meaningfully. Furthermore, a clear definition of the variables is rarely given and a variable name can have different meaning over different OGD depending on the data and the

ministry considered. A table of concordance sometimes exists within one entity, but not always. There is also the possibility that data are conceived to be used at the national or state level but researchers may need the data disaggregated at a lower level, like the district level, which requires many manipulations and assumptions.

- Insights from the workshop with researchers and official statisticians

Asher and Novosad (2015) identified challenges faced by researchers in using industry, firm, MSME and employment data in India. Most of them can be extrapolated from this specific area and remain relevant to socio-economic data sources in India, and among them OGD. They noticed the difficulty in linking records among them, even when collected by the same public entity. There is a strong need for standardisation and cooperation among official data producers. For instance in the case of microdata at the firm level, a unique identifier could be put in place and the same can be used in all similar datasets.

During the workshop's discussions many researchers expressed their dissatisfaction with the low quantity of datasets published by the public authorities compared to the amount of data collected. Lack of comprehensive data interferes in several ways with the research work. Professors from a top statistical research and training institute in India lamented the loss of statistical significance in their study due to the need of compiling different datasets coming from varied sources. This often compels researchers to make various adjustments in the data and also to make simplifying assumptions necessary to circumvent the problem of certain datasets not being made available by the government. This is particularly true for micro-level data or data related to public finance. The GoI must consider these issues in future while making data publicly available as part of OGD.

Several researchers also highlighted that more datasets on many sectors and transversal questions could be collected alongside major data collection exercise regularly organised by the government, through a small add-on questionnaire for instance. These datasets could be disclosed only on the official OGD portal and be a source of differentiation from other OGD sources. More broadly, the set-up of such portal could be complemented by the creation of transversal, inter-sectors and inter-disciplines data collection exercises and data collection teams. The selection of the data to gather could be extracted from the comments and requests posted on the portal.

During the workshop the representative from the Computer Centre of MOSPI described the initiatives undertaken over the past few years to make their data more accessible and also more usable through the preparation and publication of metadata files. Referring to the OGD official portal, the official explained that this active website is extensively used (satisfactory number of users per month) but it is only a data repository where all official agencies upload data without following-up on its usage. Moreover, this website generates many queries related to the datasets and those are sent directly to the Computer Centre. This Centre, given its current staff complement and volume of work, cannot always respond to such queries in a timely manner. The absence of troubleshooting tools and any further assistance were also generally recognised as critical issues faced by researchers to be able to use OGD successfully for their studies. He also admitted that the high turn-over within these services was inimical to the rapid implementation and success of these initiatives. More detailed insights from the workshop discussions are provided in Annex I.

Suggestions to improve the use of OGD in India

Among the challenges complicating the interaction between OGD and researchers is the lack of awareness of the OGD available. In developing countries the OGD movement is less popular, in particular among junior-level local researchers. An awareness campaign targeting the research community can be beneficial for promoting the use of OGD among them. In general, the consultation of these datasets and their use by the civil society is lower in India and in developing countries. An awareness campaign targeted all potential end-users and the civil society would also be appropriate.

Regarding data collection by official statisticians, from our research it can be said that they are aware of most of the limitations identified in this paper. For instance, officials at MOSPI indicated that a commission is in charge of improving the design and data collection system in place, including reviewing suggestions made by the data users. They are willing to make changes and improvements, especially as they see positive outcomes from the publication of data. They recognised that by publishing data the visibility of the institution increases. Official statisticians also declared that discussions in media and socio-economic journals help to give feedback on their work and decide on data collection priorities of the Ministry.

Despite limited resources and capacities, MOSPI admitted the necessity to invest in such capacities and infrastructure to bring their data dissemination system to the next level. Researchers from the civil society are willing to participate fully in this effort. Consideration was given to how researchers could be involved in the reinforcement of these capacities. For instance suggestions were made to develop short internship programs to improve dissemination practices. Official statisticians have never denied the fact that many datasets were kept unpublished, mainly due to the lack of necessary resources. Within public agencies and even within one entity, it is crucial to improve cooperation and standardisation among data producers and data collection exercises.

However, the Indian OGD portal appears to lack the possibility of establishing a real interaction among data users and data producers. This was a common conclusion from both sides, many official statisticians admitting that they did not know much about the data use by the researchers and that they did not have the structure or the resources to respond to the questions and feedback sent by the researchers about the data.

To respond to the problems generated by inappropriate data dissemination formats, the major one being the time taken to make them amenable to any kind of scientific analysis, datasets should be published in ready-to-use formats, like CSV files, XML etc. and must be accompanied by at least a brief description of the variables, data collection methodology, sampling design, etc. Moreover, the data should be disseminated online, without any charges, to conform to the OGD policies followed worldwide. All public portals should have a user registration process to gather useful information on data users and data use, and should also include an effective assistance tool and a medium to interact with data specialists. Due to the public resource scarcity, a third party, such as a data intermediary organisation, could host a wiki or other documentation portal. They would act as an intermediary between researchers with questions and data producers and official statisticians with answers.

Asher and Novosad (2015) made a series of suggestions in their paper, applicable to all OGD. They suggest following best practices in the use of information technology in data collection.

For instance tablets and other electronic devices should be used for data collection, as was the case with the 2011 Socio-Economic and Caste Census. This provides an immediate opportunity to improve data quality: 'time-stamps and geo-coordinates ensure that enumerators are actually moving from location to location and conducting surveys in a realistic time frame' (p15). The authors also suggested setting up a data extraction centre possibly through a third party (with secured machines, and export results tables, but without the possibility to download confidential data). This system has been implemented successfully in other countries and provides access to required data after removal of confidential information, and will be of great interest for researchers.

Finally we suggest, in line with most researchers consulted that data access should be made more available and free of charge as the public authorities have more to gain from an expansion of the social-economic research being conducted about India and from an increased cooperation from researchers than from a misinterpretation and misuse of data. More datasets need to be released among those already collected, with solutions to be found for any legislative or confidentiality issue. More datasets should also be collected, on specific points not yet covered by major data collection exercises or on transversal topics which study is limited due to the difficulties identified in combining datasets.

Regarding the official OGD portal and its efficient use, it should include all datasets easily downloadable, and not just a link to further sources. One recommendation to the public authorities and official statisticians to increase the traffic and use of the data displayed on this website would be to conduct a dedicated study among their selected target group to identify the key features expected. An interesting direction for the official OGD website would be to offer unique data dissemination and representation tools. This would also allow civil society to understand the data better and to contribute to the objective of transparency and accountability.

Conclusion

The study provided an overview of the literature on the broad topic of evidence-based policy-making and OGD. It highlighted the main barriers to research affecting policy, as well as the need for a more efficient research–policy connection. It further elaborated on the various possibilities arising with the use of OGD for impacting policy-making. The OGD initiative is well set up in India with a strong legal framework and support from the public authorities. However, the movement remains confined among local researchers and could benefit more if sufficient awareness and outreach efforts are undertaken. Researchers aware of OGD were generally using open public data long before the launch of the specific OGD portal. If they value its accessibility they expect more from OGD: more data released, in a better and more user-friendly format and accompanied by relevant metadata. Those elements would help researchers to make better use of OGD in their work.

Policy-makers are generally not aware of the research studies conducted using the data they have produced. More interactions between data users and data producers are desired for better dissemination of findings which may influence public policies. Further in-depth reviews of the literature are warranted to identify knowledge gaps in defined areas highlighted in the literature review.

An online survey conducted among researchers in our extended network highlighted several insights about their use and knowledge of OGD. Among these insights, is that researchers are not aware of data that could be valuable to carrying out their work. It could be said that OGD movement in India is not being used by the Indian citizens to its fullest potential. The main impediments to use are the lack of awareness and the non-availability of metadata.

The study also highlighted that researchers who do make use of OGN in India, often struggle to understand the dataset being made available. No additional supporting information is provided that describes the dataset or its underlying sampling design and it is often difficult to reach out to official statisticians with queries to get answers in a timely manner.

The issues discussed above need to be considered in the next steps of the OGD movement in India and possibly in other developing countries, and new policies should be designed to fulfil these gaps and overcome these challenges. Some of the recommended improvements are: (1) Building better links between government data providers and researchers by involving them throughout the data collection and distribution process. (2) Improve the infrastructure and access to data as well as involve those who can benefit the most from the use of OGD, such as NGOs, practitioners and research who can make an impact on policy-making.

References

Alvesson, M., Sköldberg, K. (2009). *Reflexive Methodology: New vistas for qualitative research*. London: Sage.

Anderson, R. G., Greene, W. H., McCullough, B. D., Vinod, H. D. (2005). *The role of data and program code archives in the future of economic research*. Working Paper 2005-014C. Federal Reserve Bank of St. Louis, St. Louis, MO.

Andreoli-Verbasch, A., Mueller-Langer, F. (2014). Open access to data: An ideal professed but not practised. *Research Policy*, 43, 1621-1633.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., Wouters, P. (2004). Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 3.

Asher, S., Novosad, P. (2015). *The use of firm data for development research around the world: Implications for India*. Paper presented at the Workshop on Industry, Firms and SME Data, International Growth Centre India Central, New Delhi, 27 March.

Carden, F. (2009). *Knowledge to Policy: Making the most of development research*. New Delhi: Sage & the International Development Research Centre.

Chattapadhyay, S. (2014). *Opening government data through mediation: Exploring the roles, practices and strategies of data intermediary organisations in India*. Retrieved from: http://www.opendataresearch.org/sites/default/files/publications/sumandro_oddc_project_report_0.pdf

Chun, S., Shulman, S., Sandoval, R., Hovy, E. (2010). *Government 2.0: Making connections between citizens, data and government*. Information Polity, 15(1), 1.

Court, J., Young, J. (2003). *Bridging research and policy: Insights from 50 case studies*. Working Paper. Overseas Development Institute. Retrieved from: http://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/180.pdf

Davies, T., Alonso, J. M. (2013). *Researching the emerging impacts of open data in developing countries (ODDC)*. Position paper prepared for the Open Data on the Web Google Campus, Shoreditch, London, 23-24 April.

Gray, J., Darbishire, H. (2011). *Beyond access: Open government data & the right to (re)use public information*. Access Info Europe and the Open Knowledge Foundation.

Greenberg, D. H., Mandell, M. B. (1991). Research utilization in policymaking: A tale of two series (of social experiments). *Journal of Policy Analysis and Management*, 10 (4), 633–56.

Guerry, A.-M. (1833). *Essai sur la statistique morale de la France*. Crochard, Paris.

Hamermesh, D. S. (2007). Viewpoint: Replication in economics. *Canadian Journal of Economics*, 40, 715-33.

Janssen, M., Charalabidis, Y., Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management* (ISM), 29(4), 258-268.

Lavis, J. N., Guindon, G. E., Cameron, D., Boupha, B., Dejman, M., Osei, E. J. A., (2010). Bridging the gaps between research, policy and practice in low- and middle-income countries: a survey of researchers. *CMAJ: Canadian Medical Association Journal*, 182(9), E350–E361. http://doi.org/10.1503/cmaj.081164

Manyika, J., Chui, M., Groves P., Farrell D., Kuiken, S., Doshi, E. (2013) *Open data: Unlocking innovation and performance with liquid information*. McKinsey & Company.

Newman, K., Capillo, A., Famurewa, A., Nath, C., Siyanbola, W. (2013). *What is the evidence on evidence-informed policymaking? Lessons from the International Conference on Evidence-Informed Policy Making*. Oxford: INASP. Retrieved from: http://www.inasp.info/uploads/filer_public/2013/04/22/what_isthe_evidenceon_eipm.pdf

Nutley, S. M., Walter, I., Davies, H. T. O. (2007). *Using Evidence: How Research Can Inform Public Services*. Bristol: The Policy Press.

Open Data in Developing Countries Report (2013). *Researching the emerging impacts of open data: ODDC conceptual framework*. Retrieved from http://www.opendataresearch.org/content/2014/667/researching-emerging-impacts-open-data-oddc-conceptual-framework

Open Knowledge Foundation (2012). *Open Data Handbook*. Retrieved from http://opendatahandbook.org/guide/en/

Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLOS ONE*. Retrieved from: http://journals.plos.org/plosone/article?id=10.1371 /journal.pone.0018657

Royal Society (2012). *Science as an open enterprise: Open data for open science*. London: The Royal Society. Retrieved from: https://royalsociety.org/~/media/royal_society_content/policy/projects/sape/2012-06-20-saoe.pdf

Shepherd, E. (2015). *Implications for record keepers and records management of the Open Government agenda in the UK*. UCL-DIS Research Symposium: Open data and information governance: Recordkeeping roles, 20 May, UCL, London.

Srinivasan, T. N. (2003). India's statistical system: Critiquing the report of the National Statistical Commission. *Economic and Political Weekly*, 38(4), 322-37.

Walt, G. (1994). How far does research influence policy? *European Journal of Public Health*, 4, 233-235

Webber, D. J. (1991). The distribution and use of policy knowledge in the policy process. Knowledge and Policy, 4(4), 6-35.

Whiteman, D. (1985). Reaffirming the importance of strategic use: A two-dimensional perspective on policy analysis in Congress. *Knowledge: Creation, Diffusion, Utilization*, 6, 203-24.

Wright, G., Prakash, P., Abraham, S., Shah, N. (2011). *Report on open government data in India*. Centre for Internet and Society. Retrieved from http://www.transparency-initiative.org/reports/open-government-data-study-india

Xu, G. (2012) *The benefits of open data (part II) – Impact on economic research*. Open Data Knowledge International. Retrieved from http://blog.okfn.org/2012/10/23/the-benefits-of-open-data-part-ii-impact-on-economic-research/

Yannoukakou, A., Araka, I. (2014). Access to government information: Right to information and open government data synergy. *Procedia – Social and Behavioral Sciences,* 147, 332-340.

Annex I: Takeaways from the workshop with Governments stakeholders and researchers

Challenges	Recommendations
• Denomination of the datasets can be misleading; census and survey do not design the same type of data. This can lead to misuse for research. • Some data sources do also mix both methodologies like the Annual Survey of Industries. • Misclassification can be an issue. If there are some categorical changes over time, such changes in the data collected should be carefully documented.	• Researchers would need the Government to release more of the datasets they collect periodically. • The Government should also ask for external studies to complement the internal analysis. • Need to increase collaboration and dialogue between researchers and Government on data collection and research topics. Short internship programs could be created or expanded for instance to improve dissemination practices. • A platform to keep the discussion going about the data available, their characteristics and aggregate all the information available would be a needed tool.
• Cases of misreporting, Need to find an efficient and fixed way to collect data rigorously without putting too much burden on the respondents (firms or individuals) with limited resources. • Efficient self-reporting system is missing. But low corporate governance system/level in India is interfering.	• Solutions were found in other countries, for example a rotating panel could be a solution, and firms are sampled occasionally for some years. • It is critical to document how respondents interpreted the question while conducting a data collection exercise.
• Several law restrictions identified limiting the publications of data: Taxes/tariffs info cannot be released to the public by law, this is the case in many countries not only India. It does not discourage honest reporting. Firm level data can also not be released as per the Statistical Act.	• Statistical Act revision and other relevant regulations. • It could be worth thinking about what kind of data could be shared respecting all confidentiality rules to researchers.
• Confidentiality issue • Difficulties encountered to combine data from different sources, to link record over time.	• New common identifier to introduce. It would allow panel data to be generated and deeper studies to be conducted. • Unique identifiers system to be adopted by all agencies. Coding should be made mandatory by law at all stage, otherwise codes disappear on documents. • On this topic there is a consensus that the creation of a data centre could be a good initiative, if the framework put in place is respecting all confidentiality rules.
• Duplication of efforts, new models and methodologies among the different public agencies involved in data collection.	• It would be necessary to coordinate the efforts and changes made by Government entities for increased consistency through different databases. The resources being limited, a duplication of efforts and a multiplication of methods and models should be avoided. • In general OGD improvement needs reinforced State level support.
• Most of the data collected by public authorities is not released. Many sectors in the Indian economy would benefit for having more data available and more research studies implemented.	• Creation of a data centre and publication of more already collected datasets.

• Some data collection exercises in India are very big and mobilize lots of resources.	• Needed data could be collected through small surveys done randomly, or small extra questionnaire being conducted during the conduct of a bigger survey to limit the extra costs. The resource mobilization would be minimal but it can have far reaching impact on policy-making.
• Data is collected to match the Government agenda first. Public statisticians are mostly unaware of academic researchers' focus and interests.	• The Government fully recognized that opening up scope for conducting more research in these sectors would also be very conducive for knowledge enhancement in the particular sector. • For example on panel data, taking the panel perspective to get more accurate results, and panel will inform on data quality. Looking at both panel and cross sectional data would give information on the gaps on both sides. It's very interesting for Indian Government and policies, they would have the primary benefit, and secondly it will be beneficial for researchers.
• Design of data collection is being revised by some agencies already, but difficulties in finding ideal and common frame, and structural bottlenecks due to the data collection system and bureaucracy. • Changes in the data collection design are a problem for researchers if it is not properly documented in metadata. • The sampling methodology is often not adequate, for example the chosen sampling unit of the surveys.	• More internal cooperation among official statisticians, and public authorities conducting data collection exercises and between external researchers and public authorities. • Provide precise metadata along with datasets explaining data collection methodologies and designs, and changes that have taken place over time. • Other frames are being considered by public authorities (i.e. taxes submitted, size of investment, etc.) but no adequate unit has been found yet.
• Difficulties to access data due to payment process, especially from abroad (need to provide bank draft from India based bank for example). • Data available for a fee to 'restrain' its use by non –professionals. Authorities are concerned that data could be misinterpreted and false conclusions shared with public.	• Some agencies are making needed changes: DGCIS has improved on the dissemination aspect with the creation of an online portal to disseminate authorized data. Trade data are available for download against payment of fees by credit/debit card online. • Data should be made available for free.