Access and Use of Government Data by Research and Advocacy Organisations in India: A Survey of (Potential) Open Data Ecosystem

Sumandro Chattapadhyay
Research Associate, The Sarai Programme, CSDS, Delhi, India

+91 78381 63651
mail@ajantriks.net

Paper presented at the Eighth International Conference on Theory and Practice of Electronic Governance (ICEGOV) 2014 at Guimarães, Portugal, and to be published as part of the conference proceedings by ACM Digital Library.

This is the pre-publication version of the presented paper.

ABSTRACT

The paper presents findings from a recently competed study of the practices of accessing and using government data by selected (non-governmental and non-commercial) research and advocacy organisations in India. The study takes place in the context of the Government of India adopting an open government data policy and launching an open data portal in 2012. Although, most of the organisations interacted with in this study are yet to begin substantial usage of the open data portal, they have a longer history of working with national-scale government data. The study explores the data practices of these organisations so as to evaluate the possibilities and challenges for them to act as 'open data intermediaries' – that is organisations that mediate access and use of open data by other organisations. The findings of the study provide a cross-sectoral view of the current situation of accessing and using government data in India, and briefly reflects on the future strategies towards a robust open data ecosystem in India.

Categories and Subject Descriptors

K.4.1 [Computers and Society]: Public Policy Issues – Ethics, regulation.

General Terms

Management, Human Factors, Theory.

Keywords

India, Open Data, Open Data Ecosystem, Open Data Intermediary, Open Government Data.

1. OPEN DATA ECOSYSTEM AND OPEN DATA INTERMEDIARIES

The idea of an open data ecosystem is still a young and relatively under-defined one. Discussions of the open data ecosystem, and surveys of the typology of intermediaries populating that ecosystem, often tend to identify a single set of organisations (mostly government agencies) that undertakes the supply of open data, while other types of organisations perform various kinds of value-addition to the data and deliver data-based products and services [1] [2]. Rufus Pollock offers a sharp criticism of such 'one way street' imaginations of the open data ecosystem [3]. To illustrate Pollock's argument, we may think of the adjective 'open' as applying to both 'data' and 'ecosystem.' In other words, open data ecosystem needs to be conceptualised (and realised) as a network of creators and users of open data, where there is no unidirectional flow of open data. In an open data ecosystem, the data is shared by its creators and also by its users. The government agencies, in an open data ecosystem, may not only share open data but also get back in return value-added analysis, insights, and datasets from a wide-range of users of data.

In their pioneering work, John Hagel III and Jeffrey Rayport have envisioned the rise of 'infomediary' organisations that will provide data management as a service, and enable individual producers of data to categorically control the access to their data by various agents (that may provide data-based products and services against that) [4]. On the other hand, the economics of information literature has discussed in detail the characteristics and socially optimal form of 'information intermediary' organisations [5] [6]. Such an organised is defined by its information processing activities (can be both for-profit or otherwise) being determined by information needs of its clients. While both these types of organisations – infomediaries and information intermediaries – perform tasks that are crucial for a data ecosystem, they do not necessarily create multiple sources of open data (in various stages of value-addition), and hence do not automatically augment the supply of open data in the ecosystem. A key feature of an open ecosystem is that it allows for resources to be available in various forms so that all the constituting members of the system may access and consume it in the form

appropriate for the member concerned. These users often require data at different scales of granularity and expanse (spatial and/ temporal) that enable situational use driven by the context of final users. Availability of such data requires all or at least a diverse range of actors in the ecosystem that publish primary and/or value-added data.

In this study, the term 'open data intermediary' has been used to refer to such organisations. This term has been chosen over 'infomediary' and 'information intermediary' to emphasise that these intermediaries must not only mediate access to information for other organisations, but also mediate access to open data. Further, it is important that the provision of open data is not only driven by demands of pre-identified client groups. These open data intermediaries are, thus, expected to enhance the quality and amplify the circulation of data opened up by the government agencies, through acts of sanitising, organising, compiling, formatting, and documenting available open government data. These organisations may also additionally function as repositories of open data sourced from non-government actors. Thus, within the overall ecosystem, open data intermediaries will create focused – either regional or sectoral – loops of data flow and value-addition and augment the ecosystem as a whole.

2. A STUDY OF ACCESSING AND USING GOVERNMENT DATA IN INDIA

In early 2012, Government of India approved the first national policy for governing proactive disclosure of government data in digital formats [7]. This National Data Sharing and Accessibility Policy (henceforth, NDSAP) extends the mandate of the Right to Information Act, 2005 (henceforth, RTI) to establish policy and administrative support to enable informed citizenship, better decision-making and heightened transparency and accountability. The national Open Government Data Platform <http://data.gov.in/> (henceforth, OGD Platform) was launched by the middle of 2012 by the NDSAP Project Management Unit (henceforth, NDSAP-PMU), located in the National Informatics Centre, Department of Electronics and Information Technology. Till date (June 2014), the Platform has involved seventy government agencies to share data on over three thousand thematic catalogs. Recently, the portal has been upgraded

from a base of Drupal 6 to Drupal 7, and a web-based application has been introduced to directly visualise the data hosted on the portal in to share-able charts <http://data.gov.in/visualize3/>.

Against this background of NDSAP and OGD Platform, a study was undertaken during 2013-2014 by the author to document how various (non-governmental and non-commercial) research and advocacy organisations in India access, use and re-share government data, not necessarily obtained through the portal. The study was funded by and was part of the 'Exploring the Emerging Impacts of Open Data in Developing Countries' (henceforth, ODDC) research network managed by the World Wide Web Foundation and supported by International Development Research Centre, Canada. The exploration was motivated by the existence of a range of such organisations in India that mediate access to government data and information for other organisations that do not have the capacity and/or bandwidth to directly access that data. The objectives of the study included identifying organisations that already function as open data intermediaries, or have the potential to so. The study ignored commercial organisations, since during preliminary research, no commercial agencies could be located that re-share government data (as opposed to information) under open conditions. This paper is an early presentation of findings from the above mentioned study.

2.1 Conceptualising Open Data Intermediaries

In this study, open data intermediaries are conceptualised as organisations that share data for its access, consumption and re-usage (including re-sharing) by other organisations and individuals. Three further clarifications are needed here: (a) sharing of open data by such organisations can either be done on a commercial or a non-commercial basis; (b) shared data can either be primary (collected by the organisation concerned) or secondary (sourced from an external creator) in nature; and (c) the data intermediary organisation may or may not add value to the data before sharing it further. However, given the lack of (hierarchical) depth of the (government and non-government) data access and usage ecology in India, it is expected that often the same

organisation (especially the smaller organisations) are compelled to undertake multiple data-related functions internally. Hence, the study has not been limited only to organisations that purely function as 'open data intermediaries' but interacted with a wider range of organisations that perform that task of mediating access to data and information and enable their usage by other organisations. Many of the organisations (interacted with in the study) perform multiple data-related functions such as creating data, using data to inform within-organisational activities, sharing data with other organisations and citizens in general, training other organisations and individuals to use (collect, analyse, etc.) data, etc. A few of the organisations considered in the study do not share data with other organisations at the present, but have the potential – in terms of willingness and capacity – to do so.

2.2 Identifying Organisations for Study

Initially, a web-based survey was conducted, through mailing lists and professional networks, where the participating organisations were asked to self-identify their functions vis-a-vis government data, and also to indicate which other organisations they think the study should incorporate. This survey method was not successful as it did not generate sufficient responses, and among the organisations that responded many do not perform the function of 'mediation of data access' with much emphasis, or at all. Given this experience, the study proceeded by beginning to speak with organisations that are already known to mediate access to data for other organisations, and gathering further contacts along the way.

Apart from the most obvious criteria of whether the organisation shares government datasets and information in raw form (without focusing exclusively on open formats or licenses), some of the organisations were selected on the basis of their role in running the five national knowledge portals of India – India Biodiversity Portal (Strand Life Sciences), India Energy Portal (The Energy and Resources Institute), India Environment Portal (Centre for Science and Environment), India Urban Portal (National Institute of Urban Affairs), and India Water Portal (Arghyam). Such a method

clearly brings in the possibility of subjective bias of selection error. To counterweight this, all the organisations interacted with during the study were asked to refer to other organisations that should also be addressed by the study. Answers to this question largely generated references back to organisations that have already been (or planned to be) interacted with. The study comprised of engagements with 14 organisations, located across Bangalore (3), Chennai (1), Delhi (9), and Hyderabad (1). In terms of thematic focus, these organisations work across the following sectors: budget and governance expenditure (2), education (2), electoral and parliamentary transparency (2), environment (5), and urban development (3). The study also involved discussions with the NDSAP-PMU to understand the challenges faced by the agency when working with various Ministries to open up data. For lack of time, a small number of organisations were initially considered but finally not included in the study, such as Bangalore Urban Metabolism Project, Digital Green in Delhi, and Praja Foundation in Mumbai.

3. THREE KEY QUESTIONS

3.1 How is Government Data Accessed?

Government data is typically collected through websites of the Ministries and agencies concerned, or by directly requesting them from the government offices, or using RTI requests. While downloading data tables from government websites is a very common practice, hardly any organisation (in the purview of this study) mentioned downloading of data from the OGD Platform. A key reason for this is that since these organisations have been collecting data on particular topics for a long period, they are most comfortable downloading such data from their original creator's website (say, budget data from Ministry of Finance website, and rainfall data from India Meteorological Department website). Accessing government data in closed formats (that is, PDF files and HTML tables) is a very common experience, and this is a substantial barrier to converting downloaded data into usable data, especially for time-critical exercises. For example, Association for Democratic Reforms, which collects, compiles and analyses personal and financial information declared by electoral candidates (at municipal, state,

and national levels), rely on a large data entry team to convert PDF files (for each candidate) shared by the Election Commission of India into an usable dataset. Certain Ministries, especially those that have a long tradition of publishing data for analysis by external researchers, such as the Planning Commission or Reserve Bank of India, make their data available online almost completely in open or easily usable formats. Interestingly, Transparent Chennai reported that in a very few cases, they have received datasets over e-mail and in easily usable formats in response to their RTI requests. Simultaneously, collecting data tables directly from government offices is a fairly common practice, especially for organisations working with state- and local-level government agencies. Such data can come in both digital formats and hard copies. Most organisations consider all kinds of digitally available data as a variety of open data. They are aware that these data table often come with no or vague license details but do not consider that as a problem in actuality. Importantly, data and information collected through RTI requests, or shared through the OGD Platform (that is under NDSAP), or downloaded from government websites are all popularly seen as open data and information. It is also to be noted that several high-value datasets created by government agencies are sold as data products – such as rainfall data, physical and political maps, and Census data. It is not at all clear at what levels of aggregation the re-distribution of such data is allowed by the agencies concerned.

3.2. What are the Key Challenges in Accessing Government Data?

Unsurprisingly, most organisations mentioned that the foremost problem to be solved to make government data more easily accessible is making these data sets available online. While the growth of data hosted by the OGD Platform has been very impressive, a great wealth of data gathered by government agencies in India is not only unavailable on the Platform, but more crucially, is not made publicly accessible (either commercially or otehrwise) in digital formats at all. This challenge needs be explored much more intensively as this is a question of rethinking and re-engineering the life-cycle of government data across agencies. The reporting structure between local, state and central government agencies also complicates this issue. Many of such unpublished data

sets are not kept out of public circulation due to any specific characteristics of the data itself, such as the data having personally identifiable information, or the data containing potentially sensitive information (from perspectives of either national security or social harmony). The prevailing reason for the non-publication of such data is often a simple lack of precedence of that data being shared by the government agency, or lack of confidence of the agency regarding the motivation of a non-government organisation or individual's interest in accessing that data. Several organisations reported that building long-term working relationships and trust with government agencies is fundamental to get access to these data sets. Conversely, organisations like the ASER Centre face another kind of problem as government agencies do not collect certain data sets to begin with (such as, qualitative data for primary education). This necessitates ASER Centre to undertake collection of primary data by itself. The Energy Research Institute (TERI) also highlighted the problem of lacking regularity and completeness of government data. Overall there is a general feeling that the government has failed to revise and expand its statistical machinery, especially in the face of new technologies of collecting and managing data and increasing demand for government data from policy researchers and development practitioners. Several organisations interacted with in this study are actively involved in bridging such data gaps – either through collection of primary data, or collating data sets from various (public and private) sources, or sanitising data published by government agencies, or sharing analysed data with media houses and citizens' groups. Not all such organisations, however, embrace the open data agenda fully and adopt a wait-and-watch stance towards it. This is creating insufficiency and sectoral-imbalance in demand for open government data from non-governmental sphere, as well as reducing mediated access (through re-sharing) to open government data.

3.3. How is Government Data Shared (by Research and Advocacy Organisations)?

Majority of the organisations interacted with in this study share government data in the form of various data products – analytical briefs, detailed reports, infographics for print

media, online visualisations, and printed materials shared with various user groups. However, the sanitised, reorganised version of the data (done by the agency concerned) is shared in disaggregated form by very few of these agencies. For organisations working in certain sectors, such as analysis of budget and governance expenditure data, re-sharing of data is much lesser a concern since the original data (published by government) is often in a good, directly usable format; and the challenging task is not sanitisation of data per se but its analysis. Organisations working in sectors where official data is produced by multiple government agencies and are not published in an uniform and easily-accessible manner, such as India Water Portal and Karnataka Learning Partnership, it becomes crucial to not only share the analytical findings from the government data but also the collated and sanitised data itself. Again for organisations like ASER Centre and India Biodiversity Portal – which respectively collect and share data on quality of primary education and various government schemes, and locational and taxonomic data about flora and fauna – data sharing is part of their core activities, since the government itself produce little or no (publicly available) data on the topics. Organisations whose work involve collection of substantial amount of narrative and information responses from the government through RTI requests face technical challenges in re-sharing such information and data in accessible digital formats, as those responses are usually provided in printed form. These organisations sometimes scan and upload entire paper documents for public access but this is clearly not the most desirable solution. For organisations that tend to not re-share the raw re-organised data, or share the re-organised data only in an aggregated form, the commonly offered reasons are: (a) lack of experienced demand for raw data from researchers, media persons, and other individuals and organisations, (b) lack of confidence regarding the capability of re-users to correctly interpret the data, and also in their motivations, (c) lack of an organisational history of re-sharing data, and the systemic difficulties in creating that culture. It should be noted that even organisations that do not publicly re-share data, are most likely to share it on a case-by-case basis with various academic researchers and policy analysts.

4. TOWARDS AN OPEN DATA ECOSYSTEM IN INDIA

While this study looks at the flow of government data once it comes out of the domain of government, Neeta Verma and M.P.Gupta have done a detailed survey of the inter-agency interactions and transactions for opening up data within the government, and have listed six major challenges [8]. Two further concerns emerges out of the present study that re-emphasise and augment that list of challenges: (a) critical need for government agencies, across the bureaucratic hierarchy, to start internal usage of the data collected or managed by them, and (b) mutual difficulties created by the lack of direct interactions between government agencies that collect and manage data, and the non-governmental organisations and individuals that want to use such data. The first concern leads to treatment of government data by its original creators as something that only needs to shared with higher-order agencies but not something that is directly beneficial to the creator's own activities. This reinforces sub-optimal and mis-guided data collection practices at all levels of the government, especially the field offices. Many problems with government data cited by non-governmental organisations interacted with in this study are essentially created by lack of capacity of higher-level agencies to ensure data quality, and lack of incentive for lower-level agencies to produce reliable data. Further, the Management Information Systems used by government agencies to archive government data are designed for storage and generation of aggregated reports, and not for sharing data in an anonymised form. As government agencies often do not work with disaggregated data themselves, it becomes challenging for them to share it.

The second concern foregrounds the need for the government agencies that produce and share data to understand and respond to data needs of the non-government organisations (including commercial ones) and opportunities and risks associated with that. The lack of interactions between government agencies and non-government users lead to misunderstandings between the two in terms of their respective motivations and activities. As the experiences of Association for Democratic Reforms, Karnataka Learning Partnership and Transparent Chennai exemplify, a long-term relationship

with government agencies can translate into very effective models of data sharing. The pitfall of such agency-to-agency channels of opening up government data, however, is that the actors on both sides of that relationship automatically gain a 'gatekeeper' status, even if the organisation consciously avoids acting on such priviledges. Interestingly, the lack of direct interactions with potential re-users of data impedes data sharing practices of non-governmental organisations, as mentioned above, as much as those of government agencies. The critical challenge, hence, is to organise data users interested in accessing data created and shared by both government and non-government organisations. Unless organised into a community, the various users of open government data will continue to face its current challenges – lack of a common forum for discussing data availability, standards and quality issues with the government data; lack of augmentation of government-published data through sharing of value-added disaggregated data by non-governmental organisations; and lack of an understanding of mutual data needs and competences that continue to reinforce the existing divides within these organisations and the government. As discussed in the introduction, to create a truly open data ecosystem the sharing and re-sharing of data cannot only be an onus of the government. The challenge and responsibility of creating a culture of opening up government and non-government data exist both within and outside government agencies.

To return to the term we started with, 'open data intermediary' organisations are rare in India. The objective of the study, however, was not to locate such organisations in India, but to understand what prevents them from existing, what are the implications of their absence, and how can they be created. These early findings from the study foreground that the lack of collective planning and action by organisations interested in re-using open government data across sectors create under-representation, and sometimes even mis-representation, of the data-user community not only vis-a-vis the government agencies, but also from the perspective of those non-governmental organisations that are in a position to and have the willingness to share data. The lack of collaborative initiatives among these organisations, hence, both impedes formation of 'open data intermediaries' and shows the path toward building an open data ecosystem in India.

5. ACKNOWLEDGMENTS

I am grateful to Zainab Bawa, Michael Gurstein, Tim Davies, Nisha Thompson, and three anonymous reviewers for productive discussions and feedbacks. It was also critically enriched by insights from the members of the NDSAP-PMU. The study was supported by the ODDC network, World Wide Web Foundation and IDRC, Canada. Remaining shortcomings are mine alone.

6. REFERENCES

[1] Magalhaes, Gustavo, Catarina Roseira, and Sharon Strover. 2013. Open Government Data Intermediaries: A Terminology Framework. In Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance (ICEGOV '13) Tomasz Janowski, Jeanne Holm, and Elsa Estevez (Eds.). ACM, New York, NY, USA. DOI= 978-1-4503-2456-4/00/0010.

[2] Deloitte Analytics. 2012. Open Growth: Stimulating Demand for Open Data in the UK. Briefing Note. Retrieved from: http://www.deloitte.com/assets/Dcom-UnitedKingdom/Local%20Assets/Documents/Market%20insights/Deloitte%20Analytics/uk-da-open-growth.pdf.

[3] Pollock, Rufus. 2011. Building the (Open) Data Ecosystem. Open Knowledge Foundation Blog. March 31. Retrieved from: http://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem/.

[4] Hagel III, John and Jeffrey F. Rayport. 1997. The Coming Battle for Customer Information. Harvard Business Review. January-February. Pp. 6-11.

[5] Rose, F. 1999. The Economics, Concept, and Design of Information Intermediaries: A Theoretic Approach. Heidelberg, Germany: Physica-Verlag.

[6] Womack, Ryan. 2002. Information Intermediaries and Optimal Information Distribution. Library and Information Science Research. 24. Pp. 129-155.

[7] Department of Science and Technology, Government of India. 2012. National Data Sharing and Accessibility Policy. The Gazette of India. March 17. Pp. 74-99. Retrieved from: http://data.gov.in/sites/default/files/NDSAP.pdf.

[8] Verma, Neeta and M. P. Gupta. 2013. Open Government Data: Beyond Policy & Portal, a Study in Indian context. In Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance (ICEGOV '13) Tomasz Janowski, Jeanne Holm, and Elsa Estevez (Eds.). ACM, New York, NY, USA. DOI = 10.1145/2591888.2591949.