An Open Data Agenda for Post-2015 Sustainable Development Goals

Submission to Independent Expert Advisory Group on Data Revolution for Sustainable Development, United Nations

Written by **Sumandro Chattapadhyay** with contributions from **Tim Davies, Zacharia Chiliswa**, and **Gisele S. Craveiro.**

October 15, 2014

1. An Open Data Agenda for Post-2015 Sustainable Development Goals

Data revolution' has been one of the most remarkable categories of imagination and exploration to emerge from the report of the United Nation's High Level Panel on the Post-2015 Development Agenda [1]. The identification of availability of data on the global status of human development as a key problem area is not surprising given the experiences of measuring, monitoring and implement the Millennium Development Goals. Nonetheless, the recommendation by the High Level Panel for massive restructuring of infrastructures for generating global, reliable, comparable, and timely data is significant.

A brief note prepared by the High Level Panel explains that the 'data revolution' has two key objectives: '1) the integration of statistics into public and private sector decision making; and 2) building trust between society and state through transparency and accountability' [2]. The note also lists nine strategic interventions required to achieve these objectives. Only one of which, however, addresses the second main objective.

This submission suggests that an accountable and transparent revolution of global collection and utilisation of data for sustainable development must embrace **openness** as a fundamental pre-condition of the data concerned. In other words, a data revolution for sustainable development must be based upon global collection, usage, and publication of **open data** (relevant for purposes of sustainable development).

The Independent Expert Advisory Group on 'data revolution for sustainable development' (henceforth, IEAG) has already addressed the question of open data through defining one of its consultation areas around the concept of 'Accessible Data,' which comprises of topics related to open data, accountability, and data literacy. This submission, however, proposes that **open data** must be considered as a **cross-cutting principle and instrument of ensuring transparency, trust and security** spanning all the constituent areas of the 'data revolution' -- measuring of sustainable development goals, innovation through big data and new technologies, and addressing system challenges throughout the data landscape.

To reiterate, this submission finds the decision of the IEAG to dedicate an entire consultation area to 'Accessible Data' most encouraging and praiseworthy. However, it is necessary to simultaneously ensure that the concern for and transformative potential of open data is not contained within one consultation area within, but is discussed and deployed across the various aspects of the ‘data revolution.’

2. Embracing Open Data in ‘Data Revolution for Sustainable Development’

The discourse of ‘open data’ have emerged primarily from a ‘developed world’ context, and this has informed the efforts to define the term in terms of legal and technological considerations. The '8 Principles of Open Government Data' prepared by Tim O’Reilly, Carl Malamud and others mention qualities regarding the nature of the published data (‘complete’ and ‘primary’), how it is published (‘timely,’ ‘accessible,’ and ‘machine-processable’), and the legal status of the published data (‘non-discriminatory’, ‘non-proprietary,’ and ‘license-free’) [3]. The 'Open Definition,' largely accepted as a necessary and sufficient definition of ‘open data,’ also approaches ‘open data’ as data without any legal and technological constraints upon how it can be accessed, used, modified, and shared (for any purpose) [4]. While such emphasis upon legal and technological concerns in defining ‘open data’ is necessary, operationalisation of ‘open data’ in the global development context requires consideration of other principles that may ensure the accountability of the the data production and publication processes, and the effective and transparent usage of the published data. Some of such principles have already emerged in mainstream (‘developed world’) discussion of open data. Such as the '7 additional principles' listed by Tim O’Reilly, Carl Malamud and others that include critical issues like ‘open by default,’ ‘documentation,’ and incorporation of ‘public inputs’ [5].

This section highlights several other principles the importance of which have emerged from my own survey of research and advocacy organisations in India and their practices of accessing, using and sharing (open) government data [6] and various studies undertaken as part of the ‘Exploring the Emerging Impacts of Open Data in Developing Countries’ project of the Open Data Research Network managed by World Wide Web Foundation and supported by International Development Research Centre [7].

2.1. Openness in Data Production

2.1.1. Open Procedures of Data Collection

- The principle of **openness** must be implemented from the very **beginning** of the **data lifecycle** - from the collection (or production) of data for purposes of sustainable development.

- Open production of data for sustainable development involves implementation of **open documentation** of data collection **methods**, prioritisation of **free and open source softwares**, **hardwares** and **data formats** for data collection, as well as ensuring sufficiently detailed description of the **categories** and **assumptions** employed in the data collection process. Unless these conditions are satisfied, it will not be possible to subject the collected data to public scrutiny and will undermine accountability of the collected data.

2.1.2. Openness Regarding Potential Uses of Collected Data

- Openness in collection of data must also be implemented in terms of **explicitly communicating the reason for collection and potential usages of collected data to** the individual or collective **sources** of data (from whom the data is collected). Though not commonly included in understanding of open data, this is an absolutely critical condition in the context of interventions towards global sustainable development. The individual and collective sources of data (to be used for purposes of sustainable development) must be effectively informed of the potential usages of the collected data, as absence of this can fundamentally damage the social trust between citizens and governmental agencies.

2.1.3. Openness for Big Data from Social Networks

- The above points are especially crucial when considering **web-based social networks** as sources of collection of data to be used for informing and measuring actions towards global sustainable development. Openness in the methods, assumptions, software and hardware employed in collection of data, and prior knowledge of the those who created/shared the data regarding its potential usages, must be strictly implemented when collecting data from such networks.

2.2. Openness in Data Publication

2.2.1. Open Data as an Universal Principle for Data Revolution, or 'Open by Default'

- All **data collected** and **utilised** (by public, private, academic, or civic organisations) for **purposes** of **interventions** towards **global sustainable development** must be **published** as **open data**. This principle should cover data used both at the **international level** and also at the **national level** for such purposes.

- '**Open by default**' is a necessary (but not sufficient) condition for building trust through ensuring accountability and transparency in the strategies and techniques of 'data revolution for sustainable development.' 'Open by default' in this context emphasises treatment of the above as universal and 'normal' guiding principle for all data relevant to the processes of ensuring global sustainable development.

- Implementing 'open by default' for all data collected and utilised for the purpose of informing and/or measuring sustainable development actions will require an **expanded understanding** of **'open data**,' possibly augmenting the dominant understanding formulated by Open Definition [8]. The legal and technological understanding of ‘open data’ needs to be expanded by **(1)** an understanding of the **entire lifecycle** of the **data** concerned, as mentioned above, **(2)** carefully assessed inclusion of **existing practices** of accessing, using and re-sharing government data and information that are available publicly or in semi-public networks, **(3)** responsiveness to **different modes** of consuming, analysing and acting upon (disaggregated and aggregated) data across regions, sectors, disciplines, etc., and **(4)** an acknowledgement that **‘openness’** of data is **not** simply defined by **conditions** under which it is **published**, but **‘openness’** must also apply to its various moments of **usage** and **interpretation**. Of course, this is not an exhaustive list of concerns but an indicative one.

- Further, it is crucial to consider globally produced data for the purpose of informing and measuring sustainable development as **common resources** and not as an **asset**. This principle highlights the need for making such data **available** as open data, but also to **restrict** (and/or **regulate**) the **commercial usage** of that data in cases where commercial use **may negatively impact** one or more sustainable development goals and/or basic human rights.

2.2.2. Proactive and Reactive Disclosure of Data and Information

- **Proactive** and **reactive disclosure** of **data** and **information** must be treated as **essential complements**. While, proactive disclosure of data and information is generally categorised as ‘open data’ and reactive disclosure of the same is called ‘right to information’ or ‘freedom of intervention,’ there is a threat in treating either policies of proactive or reactive disclosure as a substitute of the other. Proactive disclosure reduces public efforts in requesting for and gathering data and information on one hand, and reduces administrative efforts in re-sharing to already disclosed data and information on the other. Reactive disclosure ensures publication of such data and information that are most valuable in the context of public interests (including those concerning accountability and transparency of governance). Both the types of policies must be enforced nationally.

- Data and information sharing by government agencies are often governed by a **complex and specific** (to the national system) **sets** of **policies**, **acts**, **rules**, and other **directives**. Policies for proactive and reactive disclosure of data and information, and governance guidelines for national portals to host the disclosed data, should effectively **cross-reference** all such relevant documents. This is critical for **avoiding intra-governmental conflicts** in publishing open government data.

2.2.3. Anonymity as a Strict Condition of Openness of Data

- Personally identifiable information in an open data set, that is information that can lead to a person (or a family) being identified from a dataset, or easily de-anonymisable open data set has the potential to cause **irreversible damage** to **human security** and **well-being**. Data to be used for the purpose of informing and/or measuring global sustainable development interventions, **under no circumstances**, should be collected, stored, analysed and shared in a form that either contains personally identifiable information or that allows easy de-anonymisation of the identities of the individuals and families mentioned in the data. It is **absolutely critical** that this is treated as a **fundamental principle** of **any data** under the **purview** of 'data revolution for sustainable development.'

- Further, **joining up** of **various data** sets referring to the **same population** of individuals or families may potentially **reveal** the **real identity(ies)** of one or more members of the population. This also poses similar threat as above. To ensure that such triangulation of real identities is not possible by combining multiple data sets, all data sets used within the purview of 'data revolution for sustainable development' must necessarily contain only **internally-consistent unique identifiers** of each individuals or families (to whom the data refer to). In other words, unique identifiers used to represent a particular individual or family in a specific data set, **under no circumstances**, may refer to the same individual or family in another data set. This principle can be only **relaxed** in cases where data is collected about the same individual or family at different points of time, and hence consistent unique identifier is needed to generate a **temporal analysis** of change.

2.2.4. National Open Data Portals

- Neither are **national open data portals** substitutes for **open data policies**. While introducing an open data portal before a fully-formed (national) open data policy can have its benefits (such as accelerating the publication and usage of open government data), national open data portal cannot be **sustained effectively** in **medium- and long-run** without an open data policy to operationalise the portal.

- This is of course not to take away focus from the critical need to build **public data infrastructures**, not only to host open data shared by various government agencies, but also providing common archival support to host open data shared by private, academic and civic organisations.

- For successful implementation of national **open data portals** it is crucial to understand the **deep backend** of such a data portal that rests across multiple government agencies and sometime conflicting claims (within government) regarding ownership and control priviledges regarding data. Establishment of a national open data portal, hence, must not be seen as building an additional public-facing (and government-facing) component of government, but as an **opportunity** to **re-configure** very **processes** of **management** of **data** across agencies.

- A government-wide incorporation of the **notion** and **practice** of open data is needed to ensure success of the national data portals. This requires re-designing the **processes** of **data production** and **circulation within** the **government**, including re-creating **incentive structures** towards publication of open, reliable and timely data.

- It is fundamental that introduction of open data practices within government is **not appropriated** as a **mechanism** of greater **surveillance** and **control** of activities of lower levels of bureaucracy by higher levels of bureaucracy. Adoption of open government data means opening up data produced at all levels of government.

- A commonly provided reason (by government agencies) for non-publication of government data is either bad quality or bad format of the data. Restructuring of government data practices towards making it more open must **emphasise** upon **publication** of all **data** that is **identified** as **'bad quality'** or **'badly formatted'** by the government agencies. Opening up of such data and establishment of feedback mechanisms are the most effective method to improve their quality.

- Further, for **citizens** and **development professionals** to **effectively participate** in global sustainable development discussions, decision-making, and actions, it is critical that they have **access** to **open government data** at the **national** level, and **comparable open data** sets at the **global level**. Also, they must, as required, be provided **technological support** to **publicly host** and **share** (local, regional or other) **primary data** collected by them to inform the global discussions and actions. Both these tasks can be ably performed by a **network** of **national** and **international** (perhaps **sectoral**) **data portals**.

2.2.5. Creating Spaces for Open Data Intermediaries

- Government-run data portals are **not sufficient** to reach out to **all** the **potential users** of **data** in the world, **nor** should they **aspire** to do so. Open data intermediary organisations - that is organisations that specialise in mediating access, use and re-sharing of open data by other organisations - must be provided space in the **national** and **international strategies** of **data revolution** to augment and amplify the supply of national and international open data, as well as its **availability** across various **media forms** - digital and analog.

- There is a **rich range** of **organisations** across the world that perform this role of **mediating** access, use, and re-sharing of government data by other organisations. While all such organisations may play productive roles in supporting other organisations and individual to work with open data, it is crucial to **ensure** that these intermediary organisations does **not stop** the **flow** (that is, **re-sharing**) of open data in the process of **offering value-added services** around open data.

- In creating institutional support and partnerships between government and open data intermediary organisations, special **support** and **capacity building** processes must be employed to ensure that organisations that have been **working** with **government data** for **long** are able to **preserve** and **build** on their **expertise** as **data-driven policy discussions** and **decision-making shift** from paper-based to web-based media .

2.3. Openness in Data Usage

2.3.1. Openness in Usage of Data for Global Sustainable Development

- Creating global trust through accountability and transparency of data-driven processes of measuring development indicators, and of designing and monitoring development interventions require **open documentation** of these **processes** and **availability** of **open data** that are **used** to drive them.

- Any **data used** for taking **local**, **regional**, **national**, and **global development decisions** must be **open data**, that is it must be **available** for **public scrutiny**. This is a **necessary** principle for ensuring **accountability** of 'data revolution.'

- Following the examples of data bank maintained by United Nations, World Bank, and other international agencies, the Sustainable Development Goal monitoring agency should create an **online data archive** to host **disaggregated** (as opposed to 'country-level averages') **data** utilised in the **processes** of measuring **Sustainable Development Goals indicators**, and informing and monitoring relevant **development interventions**.

2.3.2. Opening up Public Participation in Data Discussions

- To ensure a truly broad-based revolution in the use of data for understanding and acting upon development needs, it is necessary to **reach out** to **various communities working** with **data** and **information** in the contexts of public accountability, sustainable development, human and community rights, etc.

- While the 'data revolution' and also the 'open data' agenda focus on an imaginary of data as digital, numerical and disaggregated, it is crucial to keep in mind the multiplicity of forms in which data is accessed, used, shared and acted upon by individual citizens and various collectives. The **data revolution** will be truly **successful** if it **empowers** and **mobilises all** such **circulations** and **usages** of data, towards achieving global sustainable development. Openness of data is crucial step in enabling that, but is must be **supported** through **public consultation processes** of various kinds -- from requesting inputs for technical design (of semantic standards, technical standards, data portal architectures, etc.) decisions, to educating individuals and various groups in effectively accessing, using and re-sharing data and information to ensure public accountability and informed development interventions.

- Further and in reference to earlier discussion of 'open procedures of data collection' (#2.1.1.), to **ensure effective use** of open data by various individuals and collectives, it is **critical** to **involve** them from the very **beginning** of the **data lifecycle**, that is from its **collection**. Various forms of analysis and uses of the data require it to be available not only in **specific forms** (say, CSV or XML) but also the data structure to have **specific qualities** (say, inclusion of geo-referencing of each unit source of data). Such **usage requirements** are best **resolved** by involving actors at an early **stage** of **conceptualising** the **structure** of **data** to be collected.

- Effective usage of open data for purposes of development not only requires **'data literacy'** but also **enabling institutional contexts** and **civic skills** and **capacities** to take advantage of such institutional contexts. The success of the 'data revolution for sustainable development,' hence, depends not only on **systematic production** of global, reliable and timely data on topics of sustainable development, and **open publication** of such data to enable public, private, academic and civic agencies to participate in the global development discussion, but also on the **commitment** of the **governments** to engage in and respond to **data-driven discussion** and **decision-making processes**.

2.3.3. Opening Up Political Representation and Claim-Making through Use of Data

- The previous point **envisages** the 'data revolution for sustainable development' promoting **data-driven discussions** and **decision-making** regarding global interventions towards sustainable development. This is of course **not** to **suggest** that **all** global sustainable development related **discussions** and **decisions** should **solely** be taken on the **basis** of **available data**. What is **measured** and what is not measured is a **political decision** to begin with, and often not all members of the population have **equal abilities** or **priviledges** to be **included** or **represented** within the **collected data**. One of the key **barriers** faced by **disadvantaged groups** globally while **accessing welfare services** is the need to **belong** to a **government** and/or **official data set** that determines one's **status** as a **beneficiary** or not.

- Thus, it is crucial for a successful 'data revolution' to **faciliatate empowerment** of individuals and communities to **use** not only **existing open data**, but also to **use** (locally and globally accepted) **methods** and **techniques** of **creating open data** to enter processes of claim making and of accessing welfare services.

- To reiterate, it is **absolutely crucial** that **lack** of **ability**, **capacity**, **resources**, or **priviledges** to **represent** oneself in **data terms**, or in terms acceptable to the data-driven mechanisms of global sustainable development interventions, does **not prevent** any individual or collective from **accessing** the **benefits** of **global sustainable development** or from **bettering** one's **human wellbeing**. In other words, while the 'data revolution' initiates **struggles** to bring down **'data divides'** [9] and the **various other 'divides'** that it is determined by (and determine back), in the meantime it must **not allow individual** and **collective** to be **denied** of **development** interventions and **welfare** services for their **lack** of **capability** to **use** and **belong** to **datasets**.

2.3.4. Essential Question - Who is Empowered by Using (Opened Up) Data?

- An **essential governing question** while taking decisions regarding **conceptualisation** and **operationalisation** of the 'data revolution for sustainable development' is **who is empowered** by **collection** and **opening up** of the **specific data concerned**? The 'data revolution for sustainable development' framework may **mandate** collection, archival, usage, and open sharing of **only such data** about human, infrastructural and ecological conditions that when used will **lead** to

empowerment of people towards **greater** and **sustainable wellbeing**. This is a fundamental principle for two reasons - (1) **falling costs** of collection and archival of data at a global scale creates a **strong attraction** towards **gathering** of as much **data** as possible **without specific objectives** for their collection, and (2) availability of **open data** regarding **human wellbeing** conditions at a **global** scale has **massive commercial value**, unlocking of which may not necessarily lead to positive impacts in terms of global sustainable development goals.

- **Strict adherence** to the notion of 'open data' as data that can be accessed, used, modified, and shared for **any purpose** need to be **re-considered** especially in the context of global sustainable development. The emphasis on **any purpose** specifically **ensures** the **legal possibility** of **commercial use** of the data concerned. While any **re-consideration** of the **any purpose** principle must **allow** for **commercial activities** such development experts undertaking analysis of published data (that is, **commercial** creation of **data derivatives**), but usage of such data for **commercial decision making** purposes (such as, differential rates for agricultural loans using environmental data) can be **regulated**. One **possible legal instrument** for ensuring open data use in public interest is the **requirement** that **open data** collected and published for the purpose of **sustainable development** can only be **combined** with data itself published as **open data**.

3. References

[1] High Level Panel - The Post 2015 Development Agenda. 2013. Retrieved October 14, 2014, from http://www.post2015hlp.org/.

[2] High Level Panel on the Post-2015 Development Agenda. 2013. What is Data Revolution? Retrieved October 14, 2014, from http://www.post2015hlp.org/wp-content/uploads/2013/08/What-is-the-Data-Revolution.pdf.

[3] The Annotated 8 Principles of Open Government Data. 2013. Retrieved on October 14, 2014, from http://opengovdata.org/.

[4] Open Definition. Retrieved on October 14, 2014, from http://opendefinition.org/.

[5] See [3].

[6] Chattapadhyay, Sumandro. 2014. Opening Government Data through Mediation: Exploring Roles, Practices and Strategies of (Potential) Data Intermediary Organisations in India. Retrieved on October 14, 2014, from http://ajantriks.github.io/oddc/.

[7] Open Data Research Network. Retrieved on October 14, 2014, from http://opendataresearch.org/.

[8] See [4].

[9] Gurstein, Michael. 2011. A Data Divide? Data “Haves” and “Have Nots” and Open (Government) Data. July 11. Retrieved on October 15, 2014, from http://gurstein.wordpress.com/2011/07/11/a-data-divide-data-%E2%80%9Chaves%E2%80%9D-and-%E2%80%9Chave-nots%E2%80%9D-and-open-government-data/.

4. Profiles

Sumandro Chattapadhyay is a Research Associate with The Sarai Programme at the Centre for the Study of Developing Societies, New Delhi, India. His academic interests include history and politics of informatics in post-independence India, science and technology studies, and political economy. He is associated with DataMeet, an organisation of open data and data science enthusiasts from India, and Open Data Research Network, a community of researchers from ‘developing countries’ working on themes related to open data. He also frequently works with the Centre for Internet and Society, Bangalore, India. Website: ajantriks.net. E-mail: mail@ajantriks.net.

Tim Davies is coordinator of the Open Data Research Network for the World Wide Web Foundation, and is a PhD Candidate the University of Southampton focussing on the democratic dimensions of open data policy. Websites: http://webfoundation.org, http://www.opendataresearch.org E-mail: timdavies@webfoundation.org.

Zacharia Chiliswa is the programmes Manager at Jesuit Hakimani Centre responsible for coordinating and supporting implementation of JHC programmes and projects. His areas of interest are communication for development, open governance and public policy. Website: http://www.jesuithakimani.net/. E-mail: zchiliswa@yahoo.com.

Gisele da Silva Craveiro, is a Computer Science Bachelor, has Master degree in Computer Science and PhD in Computer Systems. Since 2005 is University of São Paulo assistant professor. Coordinates the research group Colaboratory of Development and Participation (Colab-USP). She is also in the Open Government Partnership Latin American Civil Society advisor committee. She is Open Knowledge Brazil board chairwoman. Her national and international projects and publications are focused in open budget, ranging from government transparency portals analysis, data extraction, standardization of budgetary data disclosure on the web, civic application development and open data initiatives impact research. E-mail: giselesc@usp.br.

In this note, the author does not claim to represent any of the organisations and networks he is associated with. The articulated agenda is his own, and it includes ideas shared by the three contributors -- Tim Davies, Zacharia Chiliswa, and Giselse Craveiro. The author takes full responsibility for reworking and rephrasing the ideas of the contributors into the note.