MEETING REPORT

Open data in education and research*

Data play a crucial role in fundamental as well as applied research and the generation of new data is also unprecedented due to technological advances. The big data are not treated just as something big in size that can construct insights. They need to be viewed in a holistic way that can leverage alterations in policy formulation and thus have a huge impact on research, society and development. This may lead to change in power structures and become a strong pillar for sustainable development planning. Therefore, issues and challenges to manage data in a way that they can be re-used are of prime concern. The major issues in global data management will be deliberated at length during the International Conference SciDataCon 2014, to be organized by the International Council for Science – Committee on Data for Science and Technology (ICSU–CODATA) along with World Data System (WDS) at New Delhi at the invitation of the Indian National Science Academy (INSA) during 2–5 November 2014. The National Organizing Committee (NOC) has decided to organize a series of seminars/workshops across the country prior to the main international event.

A one-day meeting was organized recently at INSA as a prelude to the forthcoming international conference. The seminar was chaired by Krishan Lal (Past President INSA and CODATA) and coordinated by Usha Mujoo Munshi (Indian Institute of Public Administration, New Delhi) and Neeta Verma (National Informatics Centre (NIC), New Delhi).

The seminar was attended by invited experts in different areas who shared their views on the core issues related to data science. Representatives from research and educational institutions, Government agencies, international organizations and the corporate world attended the seminar.

The meeting started with a welcome address by Krishan Lal. While giving the genesis of CODATA, he focused on its various activities pertaining to data management for science and technology. He emphasized on the importance of free access of data, and accessibility issues, and also discussed about quality of data. Rajendra Prasad (INSA) in his inaugural remarks highlighted the need for mechanisms to handle and reveal open data in the technology era. Alok Bhattacharya (CODATA National Committee) gave a brief overview of SciDataCon 2014 and invited participants to actively participate in the event. He discussed about the incentives provided to participants from the developing countries, particularly our neighbours (South Asian countries) in open data programme and emphasized on the participation of young scientists in the area of data science and policy issues. While discussing about technical aspects and analysis of open data, he mentioned that industry participation is also required for any kind of open data activity.

Pronab Sen (National Statistical Commission, Government of India) spoke on the 'Quality of data in education and research landscape'. He mentioned that quality of data primarily depends on the procedures through which they are collected. Thus procedural part of data collection has to be well structured since the manner of data collection has a large impact and influence on data. He also emphasized the need for obviating the inadequacy of data descriptors, so that they can be used in multivariate environments and disciplines. While focusing on classification of data and relation between two datasets, Sen mentioned that standard protocols should be used to facilitate interoperability and emphasized on the importance of standardization of metadata. He discussed about ethical practice and procedural part of data collection. He also mentioned that to make open data usable, it is required to share the methodology, e.g. sampling methods, a common platform for data collectors, etc.

Deepak Pental (Centre for Genetic Manipulation of Crop Plants) spoke on 'Open data in research' and focused on the difficulties in managing data and how to extract science from data by converting them into products. He discussed about the importance of following best practices in open data and the need for open source systems for research publishing. He also mentioned about the Government effort to start an open library mission in India.

S. K. Tandon (IIT Kanpur) threw light on how to handle large data in the internet era. He highlighted that, while the research paper is the main product of any university's intellectual output, the public fund is the main source of most of the research papers of universities. He emphasized the need for multi stakeholder approach for any open data activities and mentioned that universities should set up repositories of their intellectual output that needs to be well publicized. He also said that supplementary data of research paper should be opened up for consumption.

Valli Manickam (Administrative Staff College of India, Hyderabad) discussed about national data sharing and accessibility policy (NDSAP) and its benefits. She said that the objective of NDSAP is to facilitate 'Availability and access to data and information available in both human readable and machine readable form through a network all over the country in an obligatory and time bound output oriented manner, not violative of national security and policy, thereby permitting a wider use and accessibility of public data and information'. Manickam focused on the need for transparency in the policy. She also mentioned that more quality and validation policy can be amended in later stages, but now NDSAP is in its initial stage and primary focus is on sharing of data from various Government departments. She stated that currently, the decision is to restrict data portal only to the public/Government data. She also discussed about certain issues and challenges to NDSAP policy, management of research data and interaction platform between Government and researchers.

Neeta Verma focused on open data ecosystem and discussed about the open Government data initiative and its implementation strategy. She highlighted the open source based open Government platform developed by NIC, which can be used by any organization willing to

*A report on the one-day meeting on 'Open Data in Education and Research' organized at the Indian National Science Academy, New Delhi on 19 April 2014, as a prelude to SciDataCon 2014.

launch its open data initiative. Open data platform is also available over cloud in SAAS (software as a service) mode. While talking about the necessity and impact of open government data (OGD) platform, she said that transparency, accountability, citizen engagement, collaboration, better governance and innovation are the main impact of OGD platform. She also highlighted the significance of open data in research and education, and initiatives taken by different countries and international organizations. Need for incentives for researchers/scientists to open research data was also discussed in terms of availability of platform, trusted repository, data papers and data journals and, above all, an ecosystem of research institutions, policy makers, funding agencies and individual researchers, technologists which is critical to the success of open research data initiative.

The panel discussion on 'open data and innovation in education and research' was convened by Usha Mujoo Munshi. Rajiv K. Saxena (South Asian University, New Delhi) chaired the panel discussion. The panelists included Alok Bhattacharya (JNU), M. P. Gupta (IIT, Delhi), S. K. Gupta (IIT, Delhi), Debasisa Mohanty (NII, New Delhi), Harpreet Singh (ICMR, New Delhi) and Vinay Singh (Thomson Reuters). The main topics of the discussion included the following: (i) Will open research data help enhance the research landscape of India? (ii) What kind of incentive can be given to motivate researchers to open up data? (iii) Is there any need of a policy or guidelines or directive to be issued towards making it happen (open research data)? (iv) Do you perceive some risk in opening the data? How should that happen? (v) There is a need to set a data infrastructure. How should that happen?

M. P. Gupta discussed about proactive approach taken by the Government to open up the data and highlighted the requirement of infrastructure and manpower in Government departments. He said that it would be better to share data used behind research papers. Bhattacharya spoke about the necessity of clinical trial data and pointed out that although some data are available, they are not in the requisite format which can be analysed. He discussed about the need of data, and need to be educated about open data. S. K. Gupta mentioned that academicians do not have data readily available and data should be validated/quality checked for good journals. He also said that opening up data is a bold step. He stated that so far as data portals of India are concerned, their quality and validity issues can be addressed directly by opening these to the public, as public are validating and checking all the data. Mohanty said that some meagre charges could be levied for open access to sustain journals. He also emphasized on quality and data validation. Harpreet Singh spoke about the requirement of data repository and open access. He said that many unethical practices can be trapped in case of open access of data. He said that researchers are not ready to share the data, but are ready to share reports. According to him, willingness, reliability policy and incentives are the most important parameters. He also gave some examples where data can be detrimental. Vinay Singh spoke about big data and analytical tools for research papers. He also discussed about some general principles of good practices.

The other discussed points that were echoed by many participants during the panel discussion include: the need for open research data (ORD) which is essential for analysing cross-disciplinary layers as borders between disciplines blurred over the development period. In the context of open Government initiatives, ORD is urgently being sought for use in public funding of research projects in sustainable development planning. Sharing data of research and educational institutions is a reward by itself, as it maximizes use and practices of the data in providing greater returns from public investment in research. Needless to point out that sharing of ORD may become a standard in scientific practice soon and reserved attitudes, opinions towards it will be allayed through adoption of appropriate rational norms (as for example, enriched publications create a knowledge space in a trustworthy repository). To meet this need in India, a strong link with national statistical priorities is vital as it can ensure that developing ORD repositories will meet the real needs of more relevant official statistics for sustainable development planning. The World Bank representative shared the key features of its open data portal.

In addition, the discussions about new technology solutions that enable data sharing and integration, the need to discuss the social and cultural aspects of assembling data products were also deliberated.

Keeping these observations in view, some of the suggestions to issues and questions mooted in the deliberations include: (a) National Statistical Commission may consider acknowledging the inclusion of link of public funded S&T and socio-economic research data as a vital need in the official statistics. The crucial steps towards this may include initiating strong involvement of educational and research institutes, stakeholders and technical experts as partners to standardize S&T data classification and related issues. (b) Data archives may be developed simultaneously in ORD repository systems to meet some of the challenges of contemporary data explosion. (c) For public-funded S&T and socio-economic research projects, a mandatory condition in the Terms of Reference of submitting all research data (in soft form as well) along with submission of project report will help in the reduction of duplication/misuse of data for unscrupulous project teams. (d) Datasets of data repositories on the websites may be in PDF format to reduce chances of data hacking and viral attacks. Users can obtain the data in desired format through CDs at a minimum price from the concerned organization/institutes. This will help the participating institute to meet some expenses of ORD service. (e) The strengthening of interface between science and social sciences and thereby society is crucial for sustainable development. (f) Increasing public collaboration through crowd-sourcing, citizen science and other methods, the public/Government agencies should continue to expand the ways they collaborate and engage with the public.

In his concluding remarks, Krishan Lal appreciated the complex issues and thought-provoking questions of importance on big data raised and deliberated by the participants. He also thanked all the participants for their valuable contributions during the event.

Usha Mujoo Munshi*, Indian Institute of Public Administration, Mahatma Gandhi Marg, New Delhi 110 002, India; **Neeta Verma,** National Informatics Centre, Budh Vihar, New Delhi 110 086, India.
*e-mail: umunshi@gmail.com