Open Data Infrastructure for Research and Development

Neeta Verma, M. P. Gupta and Shubhadip Biswas

1 Introduction

Value of the research data is immense, and opening research data can foster innovation, new insights and discovery. Open research data related to health, energy, physics, mechanics, social science, economy, etc., enable these raw data to contribute in the respective domains far beyond the primary analysis. Raw research data are not only being used to validate the original analysis, but also it can be helpful to discover interrelated or newly defined hypotheses, especially when associated with other openly available data.

In recent past, it has become more common for researchers to publish their research data as open data. Funding agencies increasingly require the research data (and publications) resulting from funded research projects to be published open access. However, open data access is not yet standard practice in most disciplines, and there is no culture of data sharing and reusing among researchers. Even when researchers in these fields publishes their data in the repositories and archives, the data are usually difficult to find and to access. Many international bodies are promoting open access of publications and data, like the Open Knowledge Foundation [27], Open Data Institute [26], OpenAIRE [28], OAPEN [25], and Knowledge Exchange [18]. There is often no proper infrastructure for central repository or registry of data; unavailability of central repository of data is one of the major

N. Verma (✉)
National Informatics Centre, New Delhi, India
e-mail: neeta@nic.in; neeta@gov.in

M. P. Gupta
DMS, Indian Institute of Technology, Delhi, India
e-mail: mpgupta@dms.iitd.ac.in

S. Biswas
Open Government Data Project, Delhi, India
e-mail: shubhadip.biswas@live.com

U. M. Munshi and N. Verma (eds.), *Data Science Landscape*,
Studies in Big Data 38, https://doi.org/10.1007/978-981-10-7515-5_2

challenges, and using technology, this need can be addressed by setting up a central platform for participation and collaboration, which will allow all the stakeholders to interact and explore data provision and thus enhance the potential for value addition and innovation. Open data infrastructure should enable researchers to publish their data for its use and reuse. Setting up such an infrastructure is a cost and resource-intensive effort. There is a need for policy, governance, and financial support for setting up such an infrastructure and maintain it over the years.

The paper delves into various aspects of open data infrastructure (ODI) that includes:

- Policy formulation of ODI
- Fund support for ODI
- Implementation of ODI
- Metadata and data standards
- Data use license
- Open data ecosystem
- Data citation mechanism

Focus should also be on the improvement of the quality of information, the formation and establishment of open data culture, and the delivery of the tools and mechanism to use data. A technological robust infrastructure is essential which helps all the stakeholders to make sense of data and to ensure community participation. Proper collaborative framework can enable open data to go beyond the current level of data access and research and development to foster innovation and socioeconomic growth.

2 Literature Review

Open data have become very important in research domain. Estermann demanded that the open data initiative in academic sphere started around fifty years ago [9] with the publication of the first scientific journal in 1965, i.e., philosophical transactions of the royal society, it had the policy of founding concepts escorted by the evidence on which it was constructed (i.e., data) [3]. Opening research data evade the effort to rework on the same model, provide substantiation that the methodology used for the research was accurate and correctly implemented, display answerability of the researcher, and generate the prospect of new research findings [11], not sighted by the previous researcher(s) [12]. Several journals, particularly science journals, are in support of opening experimental data [33] so that they can be reused, reproduced, and authenticated. Similarly, disclosing research data are nowadays a prerequisite for data management planning and release research application's policy [6]. Few journals also demand for the consent that research data would be shared on request [13], while for some other journals it is prerequisite.

In view of the political, social and economic factors, and academic significance and prospect of its broader application, the idea of open data has started drawing attention of the researcher community, which is evident through the emergent topic in published research papers and main topic of discussion during academic conferences. These events have experienced a sharp growth in recent years in research projects.

Open Government Data (OGD) have also been drawing a rising notice and concern of both researchers and activists from various branch of learning, such as information technology, management studies, social and political sciences, and law, due to its broadly recognized prospect to create public value through thrusting economic growth and innovation, and methodical research, and by promoting openness and significant evidence-based diplomatic dialogue [7, 15, 32]. The idea of open data is strongly related to inventive capacity and metamorphic power [8].

Various research and studies highlight that the citizen is generally open to the idea of Open Government Data but have unclear concepts about how it may relate to their lives. A study conducted by the Pew Center for Internet and American Life found the citizen rarely connect Open Government Data with collective ideas [19]. That is, the citizen has limited awareness of how to retrieve and interpret data. Consequently, researchers have generally aimed on "intermediaries of open government data," i.e., download, infer, and maneuver data [20, 30]. These data experts are important in our current time when government uses data for decision making that impact the citizen's life. Yet, majority is unknown about how intermediaries of Open Government Data can materialize the citizen's benefits of open data and are able to link it with uses useful for the community [11]. Technology can enable better discovery, ease of access, and innovative and wide use of open data; many countries in the world have already set up open data portal for this purpose [35, 36]. Open datasets should be released on portals through a structured workflow. Using technological infrastructure, platform can enable better discovery, ease of access; open data could be easily downloaded in open formats or consumed via application program interfaces (APIs) [37]. Open data platforms can be an important instrument to engage with citizens and communities to develop new products and services using open datasets.

3 Key Elements of Open Data Infrastructure

3.1 Policy Formulation of Open Data Infrastructure

To build any sustainable open data infrastructure, there is need for a robust policy framework. Policy provision has strong impact on strategy formulation and implementation of any open data program. The issue of research data publication and its citation has been highlighted by many publishers and journals in their style guides, predominantly from the scientific domain. American Sociological

Association has mentioned about machine readability of data and persistent identifier of references for future access; University of Chicago Press has emphasized on scientific databases; National Library of Medicine has publicly opened part of the data on the Web; the Council of Science Editors has put their complete research databases on the Web [14]; and many other research organizations like Nature Scientific Data, GigaScience, F1000Research, and Geoscience Data Journal have initiated publishing research data in open domain [31]. Although mandating of publication of research data is an efficient strategy [22], some other strategies have also been adapted by few research bodies, "acknowledge Open Practices" badge by Center for Open Science [17], which is used by the journal Psychological Science. Digital Object Identifier mechanism can be implemented to enable accurate citation, which would enable to track usage of the data. Academic credit of data through journal would also inspire researchers to proactively deposit more data in the repository. Using this method, peer review of datasets for data journal would take less efforts and would also reduce the delay of publishing research papers. Data managers and data contributors would get credit, which was not possible earlier. So, policy and guidelines would result into better discoverability and conceptualization of data.

3.2 Fund Support for Open Data Infrastructure

A major obstacle for open data sharing is sustainable funding; most of the government bodies or research organizations do not have any provision of funding for data publishing or data preservation, and once the research is over. Funding data activities must be a fundamental part of the scientific research effort. Development projects of open data infrastructure may be realized through grant funding by collaborations between research organizations and government bodies [16]. Government bodies can fund such initiatives, and there are many research agencies who can fund and be benefitted from promoting the optimal use and reuse of data in which funds were invested. They can do this by encouraging good data practices, investing in data infrastructure, and raising open data awareness.

3.3 Implementation of Open Data Infrastructure

The life cycle of research data contains all the stages of data from data collection for a study to sharing and reuse. Research data have a longer span of life than its origination, i.e., research project. At every stage of the life cycle, researchers may enhance the access of data, and other researchers may use the data in new research projects. Life cycle of research data originates with the study concept, after finalizing the concept, data collection phase starts. After collecting and collating data, data are then analyzed to get research findings. Processed data can be stored in

some location (i.e., data repository, archive etc.), where it can then be discovered and accessed by other researchers in future. Share and access of research data lead to the reuse of data [1], and it makes a continual loop back to the data discovery and access stage, where the redistributed data are stored and shared for open access. The UK Data Archive provides an structured description of the data cycle which has been shown in Fig. 1.

Nowadays, data management is a crucial task of the management in any organization. Process of data management is regulating the generated information during a research project. Data management is an integral part of any research projects, and there is an increasing demand for scholars who can plan and implement standard data management practice for research organizations.

Implementation of open data infrastructure requires formulation of policies, regulations, planning, execution, and management programs to regulate, conserve, deliver, and boost the value of research data and information. This idea became popular with evolution of technology from sequential processing to random access processing. The involvement of different stakeholders across world makes the implementation of open data infrastructure more challenging, which can be addressed only through the proper policy and technical implementation.

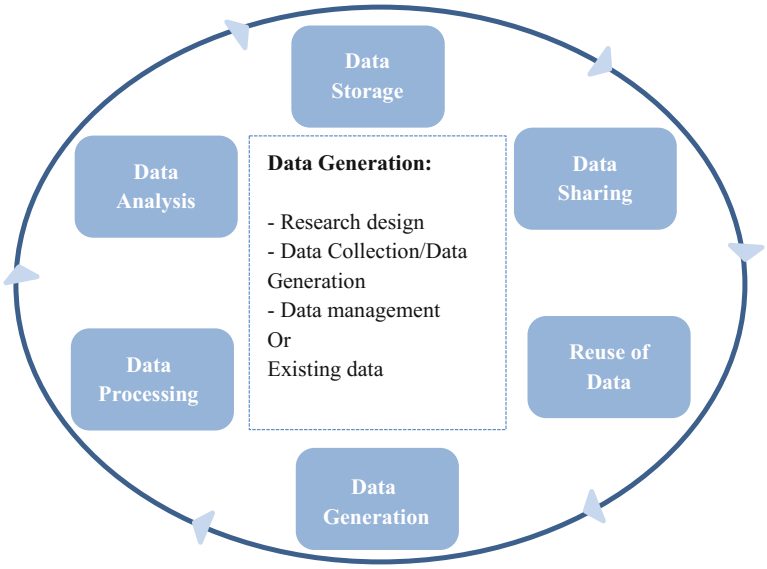

Fig. 1 Research data cycle

3.4 Metadata and Data Standards

Metadata are the key enabler to make data discoverable, utilizable, and comprehensible. To support data detection and data credentials, there are various standards, and formats of metadata have been developed over time. Metadata consist of few key elements which can be characterized according to their functionality. A standard metadata support some defined functions and describe elements to make those comprehensible [23]. Published metadata specifications with all the definitions, standards and formats should be held in a central place, and it can be published as reference file on the website or it can be kept in an accessible metadata registry.

Standardized data field is usually not followed while publishing data, as the variable names, units, and types vary across different datasets. As a result, mining of data, comparison or correlation of data is not feasible. It requires a lot of processing of data to make it ready for analysis and mashup, which is major constraint in use of data by researchers, developers, analysts, and even civil society. Hence, it is also essential that standards for metadata also be defined for open data so that the data are available in globally accepted standard format which can be used by any application confirming to those standards [34].

Better metadata or explanation of the data leads to better discovery and better reuse of data via applications or mashups. Hence, metadata are another prime aspect of data which can help data users as well as providers to use those datasets.

3.5 Data Use License

Associating an Open License with Open Data is necessary to ensure the legal grounding for its potential reuse, redistribution and is critical to ensure that such data are not misused or misinterpreted (e.g., by insisting on proper attribution), and that all users have the same and permanent right to use the data. For a data user wishing to use and build data products/services on top of the public data, they need assurance of what they legally can and cannot do with the data for both commercial and non-commercial purposes.

3.6 Open Data Ecosystem

A vibrant ecosystem can be a big influencer and can have a major role in the success of any open data initiatives. Open data can help to construct this thriving ecosystem that would create vast opportunities in research and development [5]. Open data ecosystem can be described as the provision to supply and consume open government data. Identification and involvement of the key actors would be

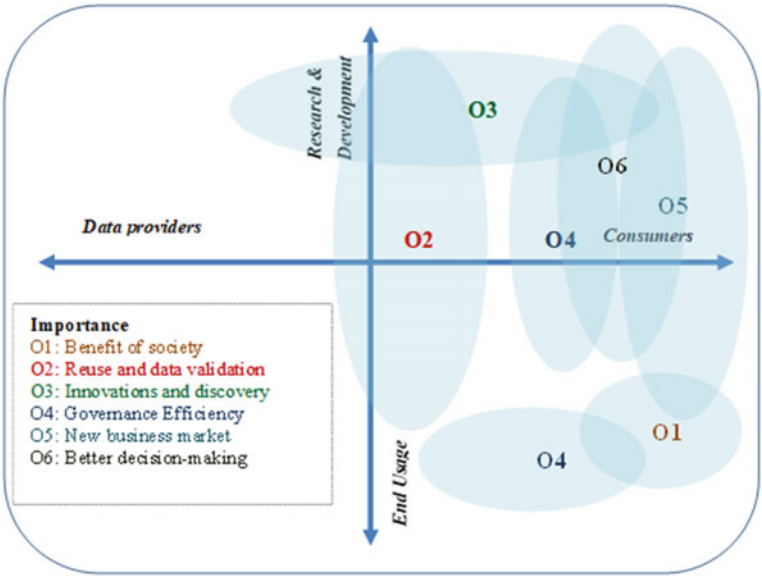

Fig. 2 Data sharing mechanism

foundation of a perfect ecosystem [20]. Policy and framework need to be developed to make the ecosystem function. A sharing mechanism has been shown in Fig. 2.

There are two sides of Open Government Data ecosystem supply side and consumption side. Data generator and providers are falling in the supply side; intermediaries and the end users will fall into consumer side.

Supply Channel. Main source of open data is governments, as they compile and generate huge data in their day-to-day operations and delivery of services. Another major source of open data is research community that may be research project or individual research. These real-world datasets require a strong data management infrastructure to supply uninterrupted data services. Therefore, data organization model would be required where data providers can share data in open format.

Consumption Channel. Consumers of open data can be divided into two major categories: One is intermediaries, and another one is end users. Intermediaries are those who will add value to the raw data shared by data providers. Application developers, researchers, journalists, data evangelists, civil society/NGOs, data scientists, and policy makers can fall into intermediate consumers of data. This group can play crucial role in creating sense of, and creating value out of, raw open data. Government, citizens, and individual users are the end users in open data ecosystem, who will use the data, products and services build out of the data. Interaction among all the players is very crucial, and understanding each stakeholder is important, as it can help to understand the potential and usage of open data [20].

3.7 Data Citation Mechanism

Another critical factor is the tracking the origin of derived data objects and primary data or scientific results from primary data collection or initial research, which may affect assurance of quality, research analysis, created model, and finally the publications. Original source is particularly important to substantiate research results used in policy formulation and decision makings, where reproducing of original experiments and procedures will not be only feasible, but also would be nearly impossible to replicate same environmental conditions [24].

Scientists are making substantial development in designing techniques to capture source information. Many scripted analysis systems built using scientific workflow models (like Kepler and Taverna) and open source data management and analytical tools [29] can be utilized to record and maintain all the data management methodology and analytical techniques which resulted into research findings. Vital information about the analytical methods can be recorded by scientific workflow applications, including details about the original data and process of its transformation, theses can provide a thorough record of an analytical model and its outputs. In this way, the analytical model, research data, and outputs turn out to be part of an information base to back the evidence-based science, to enable efficient and informed decision making in research domains [10]. New research in this domain highlights that e-Science associations are possible through the open data, and that these collaborations can be efficiently tracked to provide automatic attribution facility, i.e., where original data owner can be given credit via attributions which are resulting from provenance trace models [2, 4, 21].

4 Conclusion

The potential of open data infrastructure is enormous. Open data are a critical resource for academicians and researchers to derive a lot of insight in socio-development phenomena and in scientific research. Data are more likely to be proactively shared when the data providers and users both feel the necessity of data dissemination. There is strong sense of collaboration within research community, but they also strive for grants, for projects, for venues of publication. Researchers must interject their knowledge in the "common pot" and give up their intellectual property rights to help the collective knowledge, and they must select carefully where to apply their time and efforts. Loss of money and time can be overcome by using the open data generated by others; it would fasten the process of data collection and analysis, and will save the amount to be spent on equipment, publication fees, or other research necessities.

Sustainable release of data and maintaining open data infrastructure seems to be costly, even if the research funding is used for open data sharing, but the overall benefit may substantially decrease the cost of doing research. Data release is more

effective if the data are curated in many ways which will make it useful to others over some long period of time. Similarly, more needs to be known about potential uses and users of research data, following framework can be implemented. A central repository and open data infrastructure framework are very much required which can cater to requirements of entire ecosystem of open data. Framework can help all the stakeholders to collaborate, share and use the data; it should be designed to cater to variety of requirement of different stakeholders to enhancing the use of datasets in different ways.

Open data infrastructure must be rooted in a clear policy framework to make sure all data is accessible to all in a standard manner, while defining policy framework is important, it is also essential to establish a culture of openness. Proper collaborative infrastructure can enable open data to go beyond the current level of data availability for researchers, and community engagement could result in an unremitting dialogue between all the stakeholders; to foster innovation and socioeconomic growth, a collaborative framework of open data infrastructure has been described in Fig. 3.

The identified infrastructure and its associated elements provide a glimpse of how to implement such infrastructure, where not only the data but underlying model and research techniques would also be stored. Further research in this domain is required to refine the reference architecture and its associated elements and modules. From the process perspective, there is a strong need of a working system backed by robust policy and technological framework, where all stakeholders, i.e., researchers, governments, developers, activist, public, can join their hands to form a federated data infrastructure, not only "for the end-users," but "with the end-users." Implantation of open data infrastructure is the only way to gain

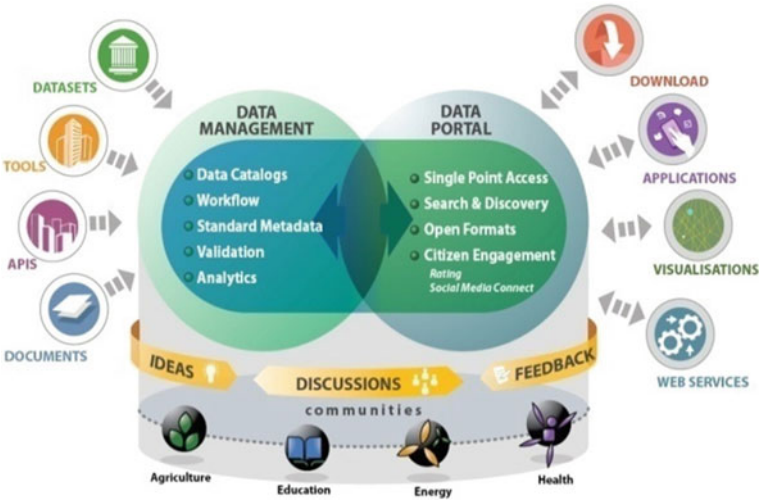

Fig. 3 Framework of open data infrastructure

confidence and acceptance of user communities, and to take the open data proposition into broader scope of innovations and socioeconomic growth.

References

1. Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). Research objects: Towards exchange and reuse of digital knowledge.
2. Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the Association for Information Science and Technology, 63*(6), 1059–1078.
3. Boulton, G. (2014). The open data imperative. *Insights, 27*(2).
4. Bowers, S., McPhillips, T., Wu, M., & Ludäscher, B. (2007, June). Project histories: Managing data provenance across collection-oriented scientific workflow runs. In *International Conference on Data Integration in the Life Sciences* (pp. 122–138). Berlin: Springer.
5. Buddenbohm, S., Cretin, N., Dijk, E., Gaiffe, B., De Jong, M., & Minel, J. L., et al. (2016). State of the art report on open access publishing of research data in the humanities. Doctoral dissertation, DARIAH.
6. Childs, S., McLeod, J., Lomas, E., & Cook, G. (2014). Opening research data: Issues and opportunities. *Records Management Journal, 24*(2), 142–162.
7. Conradie, P., & Choenni, S. (2012, October). Exploring process barriers to release public sector information in local government. In *Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance* (pp. 5–13). ACM.
8. Davies, T., & Frank, M. (2013, May). 'There's no such thing as raw data': Exploring the socio-technical life of a government dataset. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 75–78). ACM.
9. Estermann, B. (2014). Diffusion of open data and crowdsourcing among heritage institutions: Results of a pilot survey in Switzerland. *Journal of theoretical and applied electronic commerce research, 9*(3), 15–31.
10. Fox, X. M. P., Beaulieu, S. E., Fu, L., Di Stefano, M., & West, P. (2016). Documenting provenance for reproducible marine ecosystem assessment in open science. *Oceanographic and Marine Cross-Domain Data Management for Sustainable Development*, 100.
11. Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday, 16*(2).
12. Hester, J. R. (2014). Closing the data gap: Creating an open data environment. *Radiation Physics and Chemistry, 95,* 59–61.
13. Himmelreicher, R. K., & Stegmann, M. (2008). New possibilities for socio-economic research through longitudinal data from the research data centre of the German federal pension insurance (FDZ-RV). *Schmollers Jahrbuch, 128*(4), 647–660.
14. Hrynaszkiewicz, I. (2011). The need and drive for open data in biomedical publishing. *Serials, 24*(1).
15. Janssen, K. (2011). The influence of the PSI directive on open government data: An overview of recent developments. *Government Information Quarterly, 28*(4), 446–456.
16. Janssen, S., Porter, C. H., Moore, A. D., Athanasiadis, I. N., Foster, I., & Jones, J. W., et al. (2015). Towards a new generation of agricultural system models, data, and knowledge products: Building an open web-based approach to agricultural data, system modeling and decision support. AgMIP. *Towards a New Generation of Agricultural System Models, Data, and Knowledge Products*, 91.
17. Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., … & Errington, T. M. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology, 14*(5), e1002456.

18. Knowledge Exchange Homepage, http://www.knowledge-exchange.info/, last accessed 2017/08/31.
19. Lenhart, A., Simon, M., & Graziano, M. (2001). The Internet and Education: Findings of the Pew Internet & American Life Project.
20. Mishra, A., Misra, D. P., Kar, A. K., Babbar, S., & Biswas, S. (2017, November). Assessment of open government data initiative—a perception driven approach. In *Conference on e-Business, e-Services and e-Society* (pp. 159–171). Springer, Cham.
21. Missier, P., Ludäscher, B., Bowers, S., Dey, S., Sarkar, A., Shrestha, B., … & Goble, C. (2010, November). Linking multiple workflow provenance traces for interoperable collaborative science. In 5th Workshop on Workflows in Support of Large-Scale Science (WORKS) (pp. 1–8). IEEE.
22. Moore, S. (2014). *Issues in Open Research Data* (p. 164). Ubiquity Press.
23. Nogueras-Iso, J., Zarazaga-Soria, F. J., Lacasta, J., Béjar, R., & Muro-Medrano, P. R. (2004). Metadata standard interoperability: Application in the geographic information domain. *Computers, Environment and Urban Systems, 28*(6), 611–634.
24. Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … & Contestabile, M. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425.
25. OAPEN Homepage, http://www.oapen.org/home, last accessed 2017/08/31.
26. Open Data Institute Homepage, http://theodi.org/, last accessed 2017/08/31.
27. Open Knowledge Foundation Homepage, https://okfn.org/, last accessed 2017/08/31.
28. OpenAIRE Homepage, https://www.openaire.eu/, last accessed 2017/08/31.
29. Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science, 331*(6018), 703–705.
30. Sawicki, D. S., & Craig, W. J. (1996). The democratization of data: Bridging the gap for community groups. *Journal of the American Planning Association, 62*(4), 512–523.
31. Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., … & Hourclé, J. (2015). Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science, 1*, e1.
32. Stevens, B. J. (1984). *Nursing theory. Analysis, application, evaluation* (2nd ed.). Boston: Little, Brown.
33. Tananbaum, G. (2008). Adventures in open data. *Learned Publishing, 21*(2), 154–156.
34. Verma, N. (2013, August). Open data for inclusive governance. In Joint Proceedings of the Workshop on AI Problems and Approaches for Intelligent Environments and Workshop on Semantic Cities (pp. 5–5). ACM.
35. Verma, N., & Gupta, M. P. (2012). *Open government data: More than eighty formats*. Paper presented at the 9th International Conference on E-Governance (ICEG 2012), Cochin, Kerala, India.
36. Verma, N., & Gupta, M. P. (2013, October). 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* (pp. 338–341). ACM.
37. Verma, N., & Gupta, M. P. (2015, November). Challenges in publishing Open Government Data: A study in Indian context. In *Proceedings of the 2015 2nd International Conference on Electronic Governance and Open Society: Challenges in Eurasia* (pp. 1–9). ACM.