Urgent need for a data sharing platform to promote ecological research in China

Xiangzheng Deng1

Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, No. 11A, Datun Road, Anwai, Beijing, 100101 China

Abstract. China has fallen behind in data sharing. A scientific data sharing platform is needed in China to create a big data pool for large-scale comprehensive ecological research by providing undisrupted flow and sharing of data resources. To meet the increasing demand for more data-intensive ecological research, the data sharing platform should improve the quality of both data and services and handle data heterogeneity across disciplines at a higher level. As data transparency and insecurity have been major barriers to data sharing and exchange, data confidentiality and security have to be taken into consideration by governmental organizations for building a comprehensive data sharing platform. In addition, it is necessary to enhance the efficiency of data sharing among the field or laboratory ecologists in China and beyond.

Key words: big data; China; data security; data sharing; ecological research.

Manuscript received 12 April 2016; revised 25 August 2016; accepted 29 August 2016.

1 E-mail: dengxz@lreis.ac.cn

Introduction

With rapid development in ecology-related research fields, such as earth system science, environmental science, and social science, China has accumulated large amount of scientific data based on ecological observations, surveys, experiments, statistics, monitoring, and model simulations, at different levels of governmental organizations and scientific communities (Uhlir and Esanu 2006). Those data are the most fundamental and innovative resources for novel ecological research, eco-technology development, and ecological remediation in the information era. In the 21st century, as macro-ecological research is more data-intensive than ever, many ecological applications are driven by big data (Lynch 2008), which is making data sharing an irreversible trend (Dong and Li 2016). The Chinese government has recognized the importance of data sharing, such as the State Council of China issued the “Action Outline for Promoting the Big Data Development” in 2015, which emphasized the opening and sharing of governmental data and promoting the resource integration.

Big ecological data go beyond the field of ecology, as it is more efficient for stakeholders to better deal with issues based on integrated massive data resources, especially when facing gigantic emerging ecosystem issues (Provost and Fawcett 2013). For instance, the unprecedented human migrations from rural to urban areas to gain better public services, and increased transportation demands in urban areas, have not only stimulated regional economic development, but have also brought about more ecological problems across regions (Wang et al. 2016). To deal with such ecological phenomena, data sharing of multisource data and integrated social and natural analysis are required to provide policy framework and implications for environmental protection and improvement.

Scientific data sharing has been recognized as the foundation for ecological research (Peng et al. 2016). It is critical to increase data availability for results re-analysis, to provide solutions to ecological problems that require interdisciplinary interpretations based on long-term data integrity (Tenopir et al. 2011), to spur innovations exploring new hypotheses and questions with the access to the existed underlying data sets (Poldrack and Gorgolewski 2014), and to produce more cost-effective and time-saving ecological research (Tenopir et al. 2011, Ferguson et al. 2014).
Ecological research collaborations and maximize its scientific impacts at global level.

**Major Challenges**

It is quite challenging and time-consuming to formulate efficient solutions to data sharing problems due to the difficulties in data management. For example, data standardization, formatting, archiving, retrieval, visualization, and release (Robinson 2014, Peng et al. 2016) are difficult to be controlled, and data use agreements between institutions are hardly to be achieved (Tenopir et al. 2011). With rapid development of ecology-related disciplines, China has made great progress in ecological data sharing, but much more needs to be performed to meet the demand for more data-intensive ecological research. Peng et al. (2016) comprehensively elaborated the major problems that limit scientific data sharing for global change science in China from the cultural, institutional, and technological perspectives, such as data quality and heterogeneity problems, intellectual property issues, and lack of platforms and data tools. To deal with the major problems properly, many solutions have been presented, including proper archiving methods and long-term maintenance for data quality control (Thomas 2009), breaking the intellectual property barrier by rewarding co-authorship, higher citation rates (Piwowar et al. 2007), giving data sharing a greater importance in the researchers’ assessment system (Kueffer et al. 2011), and developing data visualization and access tools to permit efficient use of data (Peng et al. 2016).

**Data Quality Control**

Data quality and data services still remain important challenges to the establishment of an efficient scientific data sharing platform in China (Peng et al. 2016). The quality of data refers to the scientific basis, reliability, and effectiveness, which is subject to data providers. From the perspective of data provider, it is important to question whether data sharing is always a good thing. Data quality is difficult to monitor when the process by which data were originally collected and managed is uncertain and subjective (Tenopir et al. 2015). Where this is the case, there is no need for the data provider to engage in data sharing and it is better for users to collect new data. To rely on the data resources from different data providers, systematic mechanisms in the data sharing framework are required to evaluate the data quality and get further feedbacks from the data users to help improve the data quality. The quality of data services refers to data description and the accessibility and efficiency of data acquisition. Data description is important to improve the search accuracy and to help users appropriately apply the data. Where complete and accurate data description is lacking, it would lead users to spend more time in getting acknowledgment of the data and sometimes even to wrongly use the data. The accessibility and efficiency of data acquisition is also a big problem in many existing data sharing platforms, as some platforms provide the data download services based on confusing online systems, or based on off-line systems that may require strict application procedures. In addition, in some platforms, data are shared in a single format that does not permit access to partial data sets or permit easy data transformation. Thus, data sharing platforms should be flexible to provide parallel data processing tools that efficiently meet users’ requirements.

**Data Heterogeneity Check**

The heterogeneity of ecological data is important to be addressed for data sharing to promote ecological research. The multiple sources of heterogeneity in ecological data come from the diversities of research disciplines and the characteristics of their research funding agencies, both of which have their own terminologies, logical methodologies, experimental designs, specialized measurements, and standards (Reichman et al. 2011). Data heterogeneity in data sharing creates big challenges to the comprehensive research of single research group and also the collaborations among different research groups, as it may lead users to misuse the data and get biased research findings. The research groups therefore have to make great efforts to adopt common experimental practices, standardize their data, and regularize their methodologies. In addition, data heterogeneity arises from the diversities of spatiotemporal heterogeneity and scales. The lack of uniform standards of data processing to deal with data heterogeneity at various scales severely restricts data sharing (Liu and Zhu 2007, Ran et al. 2007, Sun and Wang 2009, Wang et al. 2013). For example, due to heterogeneity, micro-level data with regional characteristics are less likely to be used to conduct comparable research and re-analysis in different regions, nor to forecast macro-level changes. Thus, it is quite hard to implement in-depth data sharing for comparative or macro-level research over spatiotemporal scales. To deal with the problems of heterogeneity in data, strict data standardization and joint data archiving policies at a higher level of administration would be of great help.

**The Role of Government**

Government is not only a research funding agency and data user, but also an important data provider. Data confidentiality prevails among China’s governmental...
organizations at regional or national levels, and this severely limits data sharing to the public and the scientific community. In general, governmental organizations are not willing to share the data because they do not understand the value of data sharing and they are concerned over the loss of proprietary benefits. Data security is an important concern not only for governmental organizations but also for the individuals and research institutions. Although some data sharing platforms with specific regulations and rules have been set up in China, regulations and rules on data sharing in governmental organizations are lacking, and explicit data authorization mechanisms have not been adopted. As not all the data can be directly opened to all users, inappropriate data sharing may cause security concerns; thus, how to ensure the security of data archiving and transmission is a big issue to data sharing. But overly restrictive regulations would impede the reasonable use of the data resources (Liu et al. 2010). To promote the data sharing in governmental organizations, national regulations should clarify the principles, routines, and standards for governmental organizations to leverage the data resources to different type of users.

**User-Friendly Interactions**

Peng et al. (2016) have elaborated the key challenges concerning data sharing and proposed corresponding solutions from the aspects of funding agencies, research communities, and individual researchers. However, data sharing platforms must address the needs of both providers and users. From the perspective of data users, it is important to enhance the perceptions and practical experiences of the users to incorporate data sharing into their research work. In addition, the lack of interactions between different data sharing platforms and the lack of semantic-based data retrieval create data discovery problems for the data users and turn many platforms into “isolated islands.” Given the very large amounts of data shared online, users must be able to efficiently and accurately locate data sources and assess data quality. Thus, once a data sharing platform is set up, it is important to enhance the collaborations among different data sharing platforms and develop more efficient data search tools.

**Conclusions**

In summary, an efficient data sharing platform is needed to promote ecological research in China. However, China’s status of scientific data sharing is still inadequate to meet the demand for more data-intensive ecological research. Many problems and challenges for setting up an efficient scientific data sharing platform in China have been identified in previous studies. It is important to improve the quality of both data and data services. Strict data standardization and joint data archiving policies must be formulated and enforced across disciplines at high administrative levels. Increased data sharing by governmental organizations is essential. To ensure confidentiality and security, national regulations are needed to clarify the principles, routines, and standards for governmental organizations to leverage the data resources to different type of users. It is also essential to improve the perceptions and practical experience of the users of data sharing systems and to develop efficient tools for data search and application.

**Acknowledgments**

This research was supported by the National Key Research and Development Plan (Grant No. 2016YFA0602500), the National Natural Science Foundation of China for Distinguished Young Scholars (Grant No. 71225005), and the Key Project in the National Science and Technology Pillar Program of China (Grant No. 2013BAC09B00). The authors wish to thank the editors and the anonymous reviewers for their insightful contributions to this manuscript.

**Literature Cited**

Deng, R. 2016. Let scientific data sharing become the new normal for Chinese ecologists. Ecosystem Health and Sustainability 2:e01218.

Ferguson, A. R., J. L. Nielson, M. H. Cragin, A. E. Bandrowski, and M. E. Martone. 2014. Big data from small data: data-sharing in the ’long tail’ of neuroscience. Nature Neuroscience 17:1442–1447.

Kueffer, C., U. Niinemets, R. E. Drenovsky, J. Kattge, P. Milberg, H. Poorter, and I. J. Wright. 2011. Fame, glory and neglect in meta-analyses. Trends in Ecology & Evolution 26:493–494.

Liu, R. D., J. L. Sun, and B. S. Liao. 2010. Preliminary research on data licensing in scientific data sharing. Journal of Intelligence 29:15–18 (in Chinese).

Liu, R., and Y. Zhu. 2007. Explore key issues of scientific data sharing—data sharing network of Earth system science as an example. Progress in Geography 26:118–126 (in Chinese).

Lynch, C. 2008. Big data: How do your data grow? Nature 455:28–29.

Peng, C., et al. 2016. Towards a paradigm for open and free sharing of scientific data on global change science in China. Ecosystem Health and Sustainability 2:e01225.

Piwowar, H. A., R. S. Day, and D. B. Fridsma. 2007. Sharing detailed research data is associated with increased citation rate. PLoS ONE 2:e308.

Poldrack, R. A., and K. J. Gorgolewski. 2014. Making big data open: data sharing in neuroimaging. Nature Neuroscience 17:1510–1517.

Provost, F., and T. Fawcett. 2013. Data science and its relationship to big data and data-driven decision making. Big Data 1:51–59.

Ran, Y., X. Li, and J. Wang. 2007. The current key problems and potential solutions for geosciences data sharing in china. Data Science Journal 6:S250–S254.

Reichman, O. J., M. B. Jones, and M. P. Schildhauer. 2011. Challenges and opportunities of open data in ecology. Science 331:703–705.
Robinson, P. N. 2014. Genomic data sharing for translational research and diagnostics. Genome Medicine 6:1.

Sun, J., and J. Wang. 2009. Discover disperse scientific data resources sharing approach. Pages 64–74 in Integration sharing innovation—National Science and Technology infrastructure construction review and prospect. Wang, X. and L. Zhao eds. Chinese Science and Technology Press, Beijing, China (in Chinese).

Tenopir, C., S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, and M. Frame. 2011. Data sharing by scientists: practices and perceptions. PLoS ONE 6:e21101.

Tenopir, C., E. D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, and K. Dorsett. 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS ONE 10:e0134826.

Thomas, C. 2009. Biodiversity databases spread, prompting unification call. Science 324:1632–1633.

Uhlir, P. F., and J. M. Esanu. 2006. Strategies for preservation of and open access to scientific data in China: summary of a workshop. National Academies Press, Washington, D.C., USA.

Wang, Z., X. Deng, P. Wang, and J. Chen. 2016. Ecological intercorrelation in urban–rural development: an eco-city of China. Journal of Cleaner Production. http://dx.doi.org/10.1016/j.jclepro.2016.02.120, in press.

Wang, J., J. Sun, Y. Zhu, and Y. Yang. 2013. A study on the organizational architecture and standard system of the data sharing network of earth system science in China. Data Science Journal 12:91–101.

Copyright: © 2016 Deng. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.