Towards a paradigm for open and free sharing of scientific data on global change science in China

Changhui Peng,[1,2,11] **Xinzhang Song,**[3,2] **Hong Jiang,**[4] **Qiuan Zhu,**[1,2] **Huai Chen,**[5,2] **Jing M. Chen,**[6] **Peng Gong,**[7] **Chang Jie,**[8] **Wenhua Xiang,**[9] **Guirui Yu,**[10] **and Xiaolu Zhou**[2]

[1]Laboratory for Ecological Forecasting and Global Change, College of Forestry, Northwest Agriculture and Forest University, Yangling, 712100 China

[2]Department of Biology Sciences, Institute of Environmental Sciences, University of Quebec at Montreal (UQAM), Case Postale 8888, Succursale Centre-Ville, Montreal, Quebec H3C 3P8 Canada

[3]The Nurturing Station for the State Key Laboratory of Subtropical Silviculture, Zhejiang A&F University, Lin'an, 311300 China

[4]International Institute for Earth System Science, Nanjing University, Hankou Road 22, Nanjing, 210093 China

[5]Key Laboratory of Mountain Ecological Restoration and Bioresource Utilization & Ecological Restoration Biodiversity Conservation Key Laboratory of Sichuan Province, Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, 610041 China

[6]Department of Geography, University of Toronto, 100 St. George Street, Toronto, Ontario M5S 3G3 Canada

[7]Ministry of Education Key Laboratory for Earth System Modeling, Center for Earth System Science, Tsinghua University, Beijing, 100084 China

[8]College of Life Sciences, Zhejiang University, 388 Yuhangtang Road, Hangzhou, 310058 China

[9]Faculty of Life Science and Technology, Central South University of Forestry and Technology, Changsha, 410004 China

[10]Institute of Geographic Science and Natural Resources Research, Chinese Academy of Science, Beijing, 100101 China

Abstract. Despite great progress in data sharing that has been made in China in recent decades, cultural, policy, and technological challenges have prevented Chinese researchers from maximizing the availability of their data to the global change science community. To achieve full and open exchange and sharing of scientific data, Chinese research funding agencies need to recognize that preservation of, and access to, digital data are central to their mission, and must support these tasks accordingly. The Chinese government also needs to develop better mechanisms, incentives, and rewards, while scientists need to change their behavior and culture to recognize the need to maximize the usefulness of their data to society as well as to other researchers. The Chinese research community and individual researchers should think globally and act personally to promote a paradigm of open, free, and timely data sharing, and to increase the effectiveness of knowledge development.

Key words: *China; data sharing; global change science; policy.*

Introduction

Global change science, a synthetic discipline that embraces all aspects of our planet's tightly related subsystems, has become rich in data due to the rapid development of global research networks such as FLUXNET (AGNMTS; http://daac.ornl.gov/FLUXNET/fluxnet.shtml), the Global Biodiversity Information Facility (GBIF; http://www.gbif.org), the Knowledge Network for Biocomplexity (KNB; http://knb.ecoinformatics.org), the Program for Climate

Model Diagnosis and Intercomparison (PCMDI 2015; http://www-pcmdi.llnl.gov/), and the BIOME 6000 Vegetation Reconstructions (BIOME; http://www.ncdc.noaa.gov/paleo/biome6000.html). The availability of data has been expanded by increased availability of multi-sensor remote-sensing data from a large array of sensors, such as the data provided by platforms such as EOS/MODIS (http://modis.gsfc.nasa.gov) and Landsat (http://landsat.gsfc.nasa.gov), and by the long-term accumulation of data through initiatives such as the Long Term Ecological Research Network (http://www.lternet.edu). The enormous amount of available data at various spatial and temporal scales offers a tremendous opportunity to

improve simulations of Earth systems and to provide a sound scientific basis for policy development. Science is driven by information retrieved from data, and our society now relies on diverse sources of scientific data at multiple spatial and temporal scales. In particular, it is increasingly recognized that global change research requires novel strategies for open and free access to data to support policy development. It is now widely accepted that making data maximally available is a core element of scientific research, and making this the responsibility of individual scientists, research funding agencies, and the scientific community has become the operating paradigm of science in the 21st century (Hanson et al. 2011).

China is the world's most populous country and has the world's second-largest economy. Chinese government research funding has been growing at an annual rate of more than 20%, and this investment will help China to make outstanding progress in scientific and other research activities (Shi and Yao 2010). As a result of a national strategy to promote innovation through research, China's proportion of the total number of global research publications has increased greatly; China's position has improved from 17 in the world in 1993 to 2 in 2013 (ISTIC 2009; National Science Board 2014). However, the number of citations of these papers, although improving from 19 in the world in 2003 to 5 in 2013, has not kept pace with the number of publications (Fig. 1). The absence of a mechanism for open and free sharing of scientific data and various cultural considerations may be the causes of this problem because the difficulty of access to the scientific data underlying research articles has decreased their citation rate. Recent evidence suggests that sharing of research data can increase the citation rate of published papers (Piwowar et al. 2007). There is therefore a growing need to provide the global change research community with open and user-friendly access to Chinese data to fill research gaps and provide new insights into global change science.

Sharing of scientific data is a relatively new phenomenon in China. Scientific data are the basis for understanding our world and providing support for policy development and management decisions, and different policy and decision contexts lead to the production of different types of scientific information (Piwowar et al. 2007). Because global change, and particularly climate change, has become a major focus of current research, long-term collection of scientific data has increasing value for scientific and socioeconomic research. However, the value of scientific data and its benefits to both science and society can only be fully realized when these data are exchanged and shared. Data sharing makes it possible to efficiently exploit and utilize scientific data (Sun 2003), and in particular it can help researchers to avoid "reinventing the wheel." Constructing an open and user-friendly platform for sharing the available data has therefore become a key priority for developing global change and Earth system science research in China's 13th national 5-year plan.

Defining the Roots of the Problem

Traditionally, data sharing was not seen as a visible or important part of scientific research in China or a valuable scholarly endeavor. Chinese scientists have had few incentives for sharing their research data and the associated information. Rampant problems with data sharing—some attributable to cultural, institutional, and technological aspects of China's research system—have reduced the potential impact of the results from large-scale research projects for the global Earth system science community and have slowed China's potential pace of innovation. The culture of data sharing in China is still within its infancy in the field of Earth system sciences. There is a lack of general guidance within different disciplines and government funding agencies. Because of the synergies and economies that can result from this sharing, an urgent need exists to develop comprehensive data-sharing systems as well as a new cultural paradigm of open access to data. However, problems of heterogeneity and quality of data have not been fully addressed to date in China and should
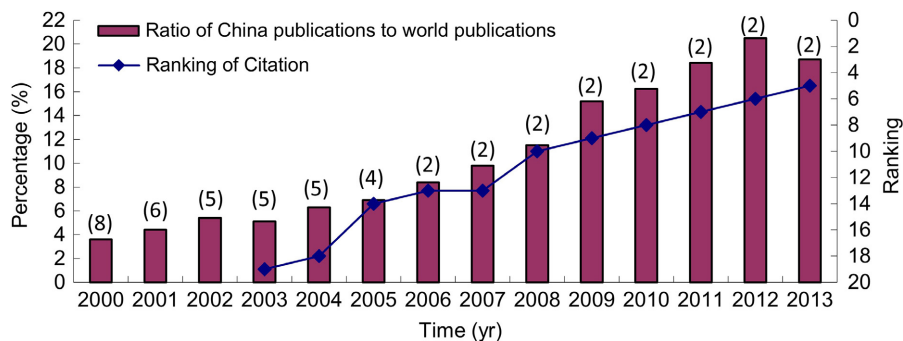

Fig. 1. The proportion of scientific publications in the world produced by China and the ranking of China in terms of citation of these publications. Numbers in parentheses refer to the ranking of the total number of publications (data source: National Science Board 2014).

be explored in depth (Reichman et al. 2011). The gross volume of scientific databases in China is still small compared with the volume in developed countries, amounting to less than 1% of the total available data volume in world databases (Pang 2009).

Cultural aspects

Sharing knowledge is the key to the progress made by science, but researchers in China do not always release their data and research materials, even after publication of their research. Most disciplines still lack the technical, institutional, and cultural frameworks required to support open data access. The idea of data sharing remains unpopular and scientists consider data collected during their research as private fruits of their labor, even though their research was funded by the government. In fact, those who undertake to share their data face many personal difficulties (Fienberg et al. 1985, Chen et al. 2014).

A major cost is time: the data must be standardized to facilitate its use by other researchers, quality control must be performed, and the resulting data must be formatted, visualized, documented, and released. Unfortunately, this investment of time is high, and most researchers underestimate how long such an effort takes. In addition to the time investment, most Chinese scientists are concerned that releasing their data may lead other researchers to challenge or contradict their conclusions, whether due to errors in the original study, a misunderstanding or misinterpretation of the data, or simply the use of more refined analysis methods. Future data miners might discover additional relationships in the data, some of which could disrupt the research agenda planned by the original investigators. Investigators may also fear that they will be deluged with requests for assistance, or will need to spend time reviewing and possibly rebutting future re-analyses of their data. In addition, they feel that sharing data may decrease their own competitive advantage, whether in terms of future publishing opportunities, exchanging information with other laboratories, or profiting from intellectual property. In some fields, however, the main barriers to data sharing are concerns about quantity and quality. Researchers must not only analyze the data and ensure that the data are correct but also need to make great effort to supply enough metadata to make the data usable by others.

As a result of these problems, most of the existing scientific data that originated from long-term investigations have not been utilized sufficiently due to the data quality issue because they do not have quality assurance and quality control (QA/QC) protocol for their data management and the absence of a culture of data sharing and mechanisms to promote this sharing. The data were therefore held by various departments and research communities as private property, making it difficult to share the data and obtain professional benefits. The US and Canada have a better data sharing and platform, policy, and system—for example, Fluxnet-Canada (http://fluxnet.ornl.gov/site_list/Network/3) and America-flux (http://fluxnet.ornl.gov/site_list/Network/1). Both provide open data (online) about continuous observations of ecosystem level exchanges of CO_2, water, energy and momentum spanning diurnal, synoptic, seasonal, and interannual time scales and are currently composed of sites from North America. However, scientists are unable to obtain the existing or published data that they need to support their research projects and must repeatedly collect similar data, at enormous cost in terms of time and money. As Shi and Yao (2010) noted, this is a problem of "wastes resources and corrupts the spirit." The intellectual property has been a key issue in China not only for research publications but also for data use and sharing. In fact, many individuals used the data of others in their publications (or reports) without even an acknowledgement. Therefore, it is critical and urgent to develop a better policy to protect the intellectual property in China.

Institutional aspects

Firstly, the mechanisms for management and coordination of data sharing at a national level are still lacking in China (Chen et al. 2014). The primary reason that data (and other resources such as equipment) have not been freely shared so far is because each institution judges the success of researchers by the number (and impact factor) of publications with either the lead authorship or corresponding authorship. Secondly, competition among institutions for research funding and awards is a real problem and big challenge for the institutional aspect of data sharing in China. Opening up the data could help competitors while reducing self-development. The absence of administrative and archival standards for data has led to scattering of data among repositories, as well as damage to, or even loss of, important data; this impedes the production of high-quality data that can be directly utilized by other scientists. The shortage of funds for database development and maintenance has made it difficult to establish fundamental public databases in some fields and to maintain them once they have been established. Another problem arises from concerns that providing full and open access to proprietary data will lead to a loss of competitive advantage due to the incomplete mechanisms for protection of intellectual property in China. The legislation that would govern sharing and utilization of scientific data has not yet been completed, so there is inadequate protection for the data owner's intellectual property. Another issue that journals and data banks face is how to ensure proper citations of these data sets. Without a generally acceptable way of assigning credit for original data

that exist outside a journal or other publication, it is no wonder that scientists are reluctant to share: their hard work may never be recognized by their employers or by granting agencies.

Technological aspects

The lack of a platform for publishing and exchanging scientific data is also a significant problem. The fact that 95% scientific databases in China do not provide an English version (i.e., most of the existing databases are only presented in Chinese) also creates a language barrier that impedes the sharing of data with researchers outside China. Most publically available databases lack data visualization tools. Such tools are vital to allow processing of complex data and to provide insights into its meaning. In addition, only a small fraction (less than 10%) of the scientific data that has been collected is accessible in a standard format or discoverable via Web searches as a result of data dispersion, heterogeneity, and provenance (Reichman et al. 2011). Most published data should be presented in standardized table formats that facilitate data mining, but developing an easily understood standard format that allows researchers to understand the meaning of the data and the relationships among the data will require the development of data definition languages and other tools. In addition, the problem of maintaining high-quality data is also vital to permit data sharing.

Although Chinese science students receive reasonably good training in statistics, their studies are rarely based on adequate knowledge of digital data technology, meta-analysis, and meta-knowledge; in addition, information management, a cross-discipline field that encompasses the entire life cycle of data, is not yet widely understood. We therefore need to find ways of educating students about non-standard data types, computational methods that can be scaled, legal protocols, data-sharing norms, and statistical tools that can take advantage of the new opportunities provided by large databases.

Recent Progress

There has been impressive progress in formulating an open data-sharing standard and developing an Earth observation database and sharing network in China in the past decade (Liu and Peng 2010, Sun 2010). In 2001, the Chinese meteorological data sharing project was launched, heralding the start of a scientific data sharing program in China. This was followed by the Scientific Data Sharing Program (SDSP) launched by China's Ministry of Science and Technology (MOST) in 2002. Its original purpose was to integrate publicly funded data resources, and its long-term goal was to leverage all possible data resources from the government to the private sector, and make this information available to the general public, thereby providing strong data to support scientific advances and innovation, government decision-making, economic growth, social development, and national security. Two years later, in 2004, the National Science and Technology Infrastructure, which included SDSP and regarded data-sharing as its core, was launched (Pang 2009). By 2010, at least 18 scientific data-sharing projects were funded by MOST in six main fields, including resources and the environment, agriculture, population and health, basic and frontier sciences, engineering and technology, and regional development.

For the first time, the Chinese Ecological Research Network (CERN) has made good progress in opening their flux tower data from 2003 to 2005 for eight key research sites across China (http://159.226.111.42/pingtai/LoginRe/opendata.jsp). In a recent meeting at Beijing (July 27–30, 2014), the CERN has collaborated and united with the US-China Carbon Consortium (USCCC), Chinese Forest Ecosystem Research Network (CFERN) for establishing a new data-sharing policy and mechanism in order to promote data sharing and data re-use. Another good example is the **G**lobal **LA**nd **S**urface **S**atellite (GLASS). It is free shared distribution of global land surface remote sensing products which facilitate the application of land surface satellite on global change research (Liang et al. 2013). Five GLASS products, including leaf area index, shortwave broadband albedo, longwave broadband emissivity, incident short radiation, and photosynthetically active radiation, have been developed, released and freely shared at Beijing Normal University (see http://glass-product.bnu.edu.cn/en/). These products have been evaluated and validated by peer-review process and they are of higher quality and accuracy than the existing products (Liang et al. 2013). Recently, China has donated to the United Nations the first open-access, high-resolution map (30-m resolution) of Earth's land cover (Chen et al. 2014). The GlobeLand30 data sets are freely available and comprise 10 types of land cover, including forests, artificial surfaces and wetlands, for the years 2000 and 2010. This will promote scientific data sharing in the fields of Earth observation and geospatial sciences (http://glc30.tianditu.com/Enbackground.html).

A series of policies and regulatory standards have been developed and documented, and institutions have been established for scientific data sharing. The need for sharing of scientific data has been added as a new section in China's Scientific and Technological Progress Law (http://english.cast.org.cn). On the basis of a survey of relevant domestic and international standards, 32 standards or regulations were designed (including 23 that are now complete at the national level). On the basis of existing standards and regulations from different industries, 120 standards and regulations for data sharing in these industries have been drafted at a division level (Pang 2009). By December 2005, the project had invested more than 250 billion RMB (*ca.* USD 30 billion).

The total users visiting scientific data-sharing websites and its sub-sites had reached 14,000,000 and registered users' number reached 50,000. By June 2008, 3,616 databases had been constructed, and the data sets that were available online amounted to more than 35.5 TB of data, of which more than 24 TB had been downloaded to support more than 1,500 research and engineering projects. Funding for these projects has amounted to more than 25 billion RMB since 2000. More than 50% of these data resources have mainly been used for scientific research and education, and especially for national key scientific research programs such as the "973" Program, the "863" Program, and the National Natural Science Foundation Program.

Chinese agencies are also making breakthroughs in international cooperation. The China World Data Center was set up in 1989. Chinese agencies have been working with US agencies to exchange data (e.g., with the USGS and NOAA) since 2005. The cooperation with the Committee on Data for Science and Technology (CODATA, http://www.codata.org/) for data sharing will be continued. More recently (by November 5, 2010), the Government of China organized and hosted the Group on Earth Observations (GEO) 2010 Ministerial Summit and produced "The GEO Beijing Declaration: Observe, Share, Inform" (GEO 2010*a,b*) to advance international cooperation on Earth observation systems, data sharing, and management in Earth system sciences. China serves as a Co-Chair of the GEO Executive Committee and has significant say in the international arena on sharing Earth observations data. One of GEO's core tasks is to develop the Global Earth Observation System of Systems (GEOSS) Data Sharing Implementation Guidelines and Action Plan (23) and to establish the operational GEOSS

Common Infrastructure, which significantly improves access to global Earth observation data and resources. The online supplemental material provides links to several key Chinese and International databases.

Future Challenges and Opportunities

Although outstanding progress has been made to improve sharing of scientific data in China, some problems related to policies, standardization, technology, and coordinated management are still impeding the progress of data sharing. Fig. 2 summarizes these problems and their relationships. Key data sharing challenges that the Chinese global change community must resolve include:

1. Data quality issue in China is rooted in lacking the standardized field data collection protocol and QA/QC procedures in place to guarantee the accuracy and precision of data. The success of the Chinese research network is dependent upon the quality of data and its accuracy. The research community requires an early and continuing commitment to the maintenance, quality assurance, documentation, and distribution of its data sets. The creation of an international compatible field data collection protocol and a long-term and high-quality data archive will be an important scientific legacy of the research networks in China.
2. A cultural shift must still occur to encourage freer data sharing, as the Chinese culture of data sharing has not yet changed dramatically. (A profound cultural change of this kind seems necessary because it remains an ongoing challenge.)
3. Data visualization and access tools must be developed to permit efficient use of data. The resulting data service
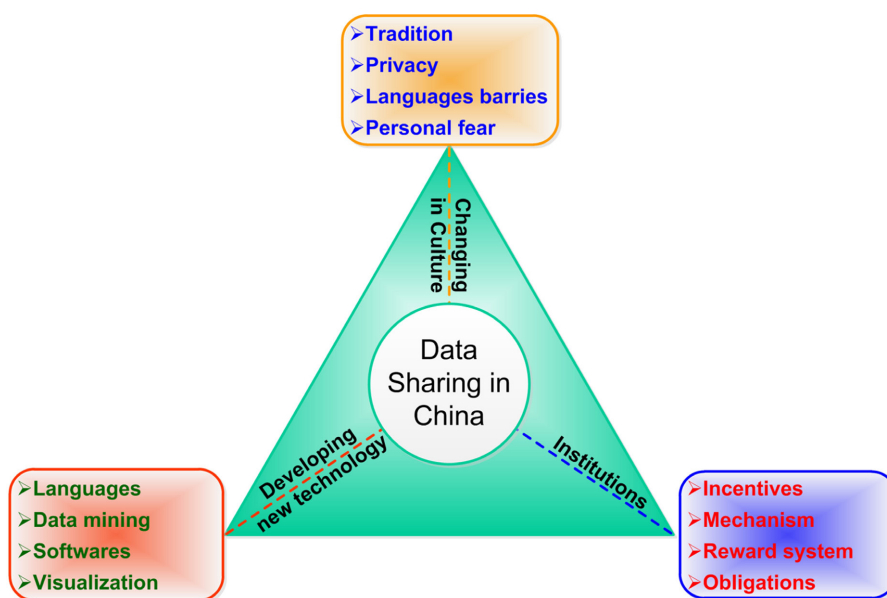

Fig. 2. Current problems (e.g., tradition, privacy, languages barriers, and personal fear) and future challenges (incentives, mechanism, reward system, and obligation) for data sharing in China.

must include assistance in selecting and obtaining data; access to data-handling and visualization tools; notification of researchers about data-related news; and technical support and referrals. New Web-based technology has already started to make Earth observation data more easily accessible.

4. It will be necessary to encourage data synthesis for global change research. China is betting that an ambitious program of data management will help to secure a future of increasingly open and free data sharing.

5. Agencies such as the National Natural Science Foundation of China (NNSFC) and MOST that fund research need to recognize that the development of both hardware and software tools for data-sharing systems is central to their mission to facilitate open and user-friendly digital data, and need to support the development of these tools with appropriate funding.

6. It will be necessary to develop a new education and outreach program to inform both experienced researchers and students about data access, re-use, and mining, as well as meta-analysis, data management, and data sharing. Data analysis and management topics should be woven into relevant courses at both the undergraduate and graduate levels.

7. Universities, research institutions, and funding agencies should develop new measures to evaluate a research project's success not only based on publications and other outcomes it produces but also based on the amount and quality of data it makes available for the wider user community and society.

8. Journals and other organizations should encourage (or demand) that researchers deposit their data where it will be publically available as a precondition for publishing. A new journal of *Earth System Science Data* (http://earth-system-science-data.net/), which is dedicated to promoting the exchange and sharing of scientific data, was launched in 2009. The journal is an international, interdisciplinary journal that focuses on the publication of articles on original research data sets, thereby promoting the reuse of high-quality data of benefit to the Earth system sciences community.

Thinking globally and acting individually

To move toward open and user-friendly data-sharing systems in China, we must both think globally and act individually and personally: funding agencies can become a key motivating force for this change if they demand data sharing in return for financial support. For data-intensive projects such as the 973 and 863 programs, a standard requirement is that all relevant data must be made available on a publicly accessible website at the time of a paper's publication. Various scientific societies can encourage this approach by establishing it as a precedent, and journals can make sharing of data a condition for publication. Domestic and international Chinese journals can learn from *Science*

and *Nature* as well as other international journals (such as the open-access Public Library of Science journals or *Ecological Archives*) to strengthen their policies to make data maximally available by publishing supporting online materials.

The emphasis on scientific data sharing will be in the following two directions. Firstly, scientific data center groups will be established for fundamental and commonweal fields, including meteorological data, marine data, and hydrological data. The goal will be to collect and re-organize all possible data from government agencies, institutes, programs, and individual investigators while making full use of international scientific data resources through increased cooperation with the international research community. The result will be a multi-tier, distributed scientific data-sharing system that bridges the gaps between agencies, institutes, and geographic regions.

Another direction will be to establish networks for sharing scientific data based on the huge amounts of data produced by major science and technology projects, key areas of research, and fundamental and frontier fields. More attention must be paid to developing an open-access culture in China and to making data increasingly accessible to all interested users at an affordable cost, or free if possible.

The investment sources and intellectual property rights related to scientific data must be strictly distinguished and an appropriate data management pattern must be established, whether by providing free access or charging a fee for utilization of the data. It is important to note that the policy should also recognize that there may be valid reasons for not sharing data, including concerns about patient privacy and informed consent in fields such as medicine. It is also understandable that some agencies may worry about how releasing some types of data might affect national security. Most participants in the development of this policy agree that some sensitive data, such as the precise locations of the last few individuals of an endangered species, should not enter the public domain. However, these instances are now being perceived as the exceptions to the rule rather than as the default assumption.

The major tasks that lie ahead will be to re-edit existing data resources; safeguard endangered scientific data and records; develop the master database for large research programs funded by the government; introduce international data resources based on their scientific value, quality, and usability; integrate data from multiple sources; and conduct value-added research.

In summary, China still lags behind the US and Europe in terms of scientific data sharing and data quality. Creating integrated data-sharing frameworks will therefore require a collective effort by all stakeholders. Several initiatives toward collaborative, community-based sharing of data have already begun, but China's sharing of scientific data is still very much at the nascent stage of overall planning, and is still accumulating experience

with the required technology and with policy development, as well as with overseeing pilot data-sharing projects. There is still a long way to go to achieve full and open exchange and sharing of scientific data. To do so, Chinese research funding agencies must recognize that preservation of, and access to, digital data are central to their mission, and must support these tasks accordingly. The Chinese government also needs to develop better mechanisms, incentives, and rewards, while scientists need to change their behavior and culture to recognize the need to maximize the usefulness of their data to society as well as to other researchers. The Chinese research community and individual researchers must think globally and act personally to promote a paradigm of open, free, and timely data sharing. Chinese universities and research institutions must also develop a new education and outreach program to teach their teachers, students, and graduates about the need for data access and sharing. It is time for China to build a healthy culture of data sharing and a platform for data sharing that is compatible with international resources.

Acknowledgments

Literature Cited

Chen, J., Y. Ban, and S. Li. 2014. Open access to Earth land-cover map. Nature 514:434.

Fienberg, S. E., et al. 1985. Sharing research data. National Academy Press, Washington, D.C., USA.

GEO. 2010*a*. The Group on Earth Observations: GEOSS Data Sharing Action Plan for the Implementation of the GEOSS Data Sharing Principles. GEOSS Strategic Targets, as accepted at GEO-VI. http://www.earthobservations.org/documents/geo_vi/12_GEOSS%20Strategic%20Targets%20Rev1.pdf

GEO. 2010*b*. Group on Earth Observations Beijing Declaration: observe, share, inform. http://www.earthobservations.org/documents/ministerial/beijing/MS1_The%20GEO%20Beijing%20Declaration.pdf.

Hanson, B., A. Sugden, and B. Alberts. 2011. Making data maximally available. Science 331:649.

ISTIC. 2009. Statistics of Chinese publications in 2008. Institute of Scientific and Technical Information of China, Beijing, China. (in Chinese).

Liang, S., X. Zhang, Z. Xiao, Q. Liu, J. Cheng, and X. Zhao. 2013. Global LAnd Surface Satellite (GLASS) Products: algorithms, validation and analysis. Springer, New York, New York, USA.

Liu, R. D., and J. Peng. 2010. Construction and prospect of scientific data sharing policy & regulation in China. Science and Technology Management Research 13:40. (in Chinese).

National Science Board. 2014. Science and engineering indicators 2014. National Science Foundation, Arlington, Virginia, USA.

Pang, Y. 2009. Preliminary study on current status about digital science sharing and S&T literature sharing at home and abroad. China Science and Technology Resource Review 41:10. (in Chinese).

PCMDI. 2015. Program for climate model diagnosis and intercomparison. http://www-pcmdi.llnl.gov/

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.

Reichman, O. J., M. B. Jones, and M. P. Schildhauer. 2011. Challenges and opportunities of open data in ecology. Science 331:703–705.

Shi, Y., and Y. Yao. 2010. China's research culture. Science 329:1138.

Sun, J. L. 2003. Science data resource and sharing. China Basic Science 1:30. (in Chinese).

Sun, J. L. 2010. It is not expected so difficult to realize data sharing. Science Times Online. 3 November 2010 (in Chinese). http://news.sciencenet.cn/dz/dznews_photo.aspx?id=10042

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1002/ehs2.1225/suppinfo