China's scientific progress hinges on access to data

It is getting harder for scientists in China to obtain the high-quality public data that they need for important research studies, says **Zheng Wan**.

The era of big data has seen the development of bioinformatics databases, data sharing and increased access to technical resources. Until recently, this trend towards open science was helping scientists in China to compete on the world stage. However, growing constraints on public-data exploration and sharing are being felt throughout Chinese society. If not addressed, they will slow scientific research and innovation.

It is hard — and getting harder — for Chinese scientists to access high-quality domestic data. Most of the public data are held by government departments, some of which are strengthening their monopoly and making it harder for researchers to access the information. This affects researchers in the humanities and social sciences especially, but also extends to fields such as environmental science and public health, because the data involved can be politically sensitive. At conferences, I hear numerous complaints from colleagues about how hard it is to extract routine figures such as air-pollutant levels from the authorities, for example.

Even when data are published, some are likely to be of poor quality because they have not been collected properly. The most notable example is the controversy on China's gross domestic product (GDP). There is a significant — and widening — difference between the official national estimate and the total calculated by adding up the GDPs of each of China's 31 province-level divisions. The National Bureau of Statistics in Beijing admits that different data-collection methods are used at the provincial level, and is trying to harmonize them. So far, progress has not been encouraging.

Public data sharing has been turned into a profit-making scheme. It would be useful, for example, to compile data on pollution from road vehicles in China. Done properly, this would require access to detailed records on the number of each type of vehicle licensed, road congestion, detailed engine parameters and fuel standards. Research institutions struggle to get even basic figures on vehicle ownership from public agencies, so they must use less rigorous — and often misleading — sales data that industry groups collect from manufacturers. Ironically, the same wealthy automobile manufacturers that inflate their own numbers can get objective, reliable data about their competitors' sales by buying them from special channels that are linked with some government departments — at a price that public institutions and scientists cannot afford.

In such an environment, it is no surprise that some research teams in China do not want to publish their own data. Ownership of data represents intangible capital that gives scientists a competitive advantage in some academic fields. My own research group receives many requests for maritime data, such

as port statistics and fleet information, that we have compiled, but we are reluctant to share the information. The workload and cost of sourcing and sorting scattered data sets into a usable form are enormous. If we keep these data exclusive, we can use them to develop research papers. If it were easier to access good data from other sources, we would be more comfortable with giving our own away.

Open access to and improved quality of public data can promote transparency in government affairs. Despite the slow progress, there are positive examples of improved transparency and how it has benefited Chinese society. A notable one is the full disclosure of air-pollution data, which started in 2014 as a result of mounting public pressure on environment bureaus. Before this, only sketchy data were published on a daily basis. These data are now updated hourly and are widely shared between government agencies in China. The data allow the health bureau to send alerts to the general public, the education bureau to decide whether classes should be suspended during smoggy days and the transportation bureau to adjust its vehicle-restriction rules. Perhaps the most important effect of the full disclosure is heightened public awareness of the worsening pollution crisis. Unfortunately, not all data that interest scientists also interest the general public. There is still a long way to go in achieving full transparency and increasing the availability of public data.

The restrictions are not confined to information generated and held inside China. Foreign academic resources can also be technically demanding to access. Several information-management bureaus have set up digital roadblocks to filter supposedly harmful information.

My life as a working scientist in China is affected. Reliable searches of the academic literature are near impossible. With no access to Google Scholar — which I prefer over other search engines, because it combines books, papers, theses, patents and technical reports — I have to keep track of trends by individually searching the databases operated by publishers that are still accessible.

Secure networks are crucial for national security, but good data are the backbone of scientific progress and economic development. Resource sharing and public access to trustworthy information underpin economic and social well-being. In China, resolving these conflicts will entail a comprehensive study, so that we can establish a highly efficient and reasonable data-management mechanism that benefits all. More immediately, researchers should be granted greater access, especially to public data and to academic search engines. ∎

> SECURE NETWORKS ARE CRUCIAL FOR NATIONAL **SECURITY,** BUT GOOD DATA ARE THE BACKBONE OF SCIENTIFIC **PROGRESS.**

Zheng Wan *is a research scientist at the College of Transport and Communications, Shanghai Maritime University, China.*
e-mail: mrwan@ucdavis.edu