OPEN ACCESS TO SCIENTIFIC DATA: PROMOTING SCIENCE AND INNOVATION

*Guan-Hua Xu**

** Minister, Ministry of Science and Technology, Beijing, the People's Republic of China.*

Email: zhangxe@most.cn

ABSTRACT

As an important part of the science and technology infrastructure platform of China, the Ministry of Science and Technology launched the Scientific Data Sharing Program in 2002. Twenty-four government agencies now participate in the Program. After five years of hard work, great progress has been achieved in the policy and legal framework, data standards, pilot projects, and international cooperation. By the end of 2005, one-third of the existing public-interest and basic scientific databases in China had been integrated and upgraded. By 2020, China is expected to build a more user-friendly scientific data management and sharing system, with 80 percent of scientific data available to the general public. In order to realize this objective, the emphases of the project are to perfect the policy and legislation system, improve the quality of data resources, expand and establish national scientific data centers, and strengthen international cooperation. It is believed that with the opening up of access to scientific data in China, the Program will play a bigger role in promoting science and national innovation.

KEYWORDS: Scientific data, Data sharing, Science policy, Research infrastructure, National innovation

1 INTRODUCTION

As economic globalization speeds up, the global flow and allocation of essential factors of productivity become much more popular than ever before, especially in capital, information, technology, and talents. The advancement of technology and innovation are becoming the main ways of improving the overall strength and core competitiveness of each nation. Reliance on science and technology to realize the sustainable use of resources and to promote the harmonious development between humans and nature is already a universal strategic goal.

In the face of many challenges and opportunities, in early 2006 the Chinese State Council released the National Guidelines for Medium- and Long-term Plans for Science and Technology Development (2006-2020). The ultimate goal is to transform China into one of the innovative countries. The Guidelines establish a target to raise the level of China's research and development expenditures in GDP to 2.5 percent or above, with a science and technology advancement contribution rate reaching 60 percent. One of the strategic elements of the Guidelines is to construct the science and technology infrastructure platform, which is fundamental to building China's capacity in scientific and technological research.

Scientific data sharing is one of the core elements in these plans. The Ministry of Science and Technology (MoST) gives high priority to scientific data sharing and therefore launched the Scientific Data Sharing Program (SDSP) in 2002.

2 THE OVERALL CONCEPT OF PROMOTING SCIENTIFIC DATA SHARING IN CHINA

Based on the principles, objectives and tasks defined by the Construction Outline of the National Science and Technology Infrastructure Platform, MoST establishes the overall approach for promoting scientific data sharing in China. The 2006 State Council Guidelines promote the integration of scientific data resources generated and accumulated by national research projects, with a focus on public welfare and basic science, to make them more open and accessible based on the requirements of scientific and technological innovation. The defining principles and priorities include overall planning and resource sharing, cooperative development of unified standards, demand-oriented activities and guaranteeing of security. Several pilot projects are being initiated in these areas.

By 2020, China is expected to achieve several key objectives through the implementation of SDSP. It will establish a networked scientific data management mechanism and a data sharing service system with an effective structure and broad coverage of most basic science and public-welfare domains. It will establish data policies, regulations and standards, and implement an operational sharing mechanism. The Program will develop a technology-oriented service team with appropriate professional representation and the ability to adapt to social needs in the application of information. It will open up access to over 80 percent of the public-welfare and basic science data resources. Overall, the Program will make the accumulation and sharing of scientific data resources support the basic requirements of innovation, and ultimately promote economic and social development.

Already by 2010, China is expected to build a data management and sharing service system with a three-tier structure consisting of 40 scientific data centers or networks, 300 master databases and one portal. This system will cover six major fields: natural resources and environment, agriculture, population and health, basic and frontier sciences, engineering and technology and regional scientific and technical research.

The SDSP was initiated to function as a catalyst, to integrate publicly-funded data resources with a view to leverage all possible data resources from the government to the private sector, and to make them available to the general public. Figure 1 illustrates the Program's three-tiered structure as outlined above.

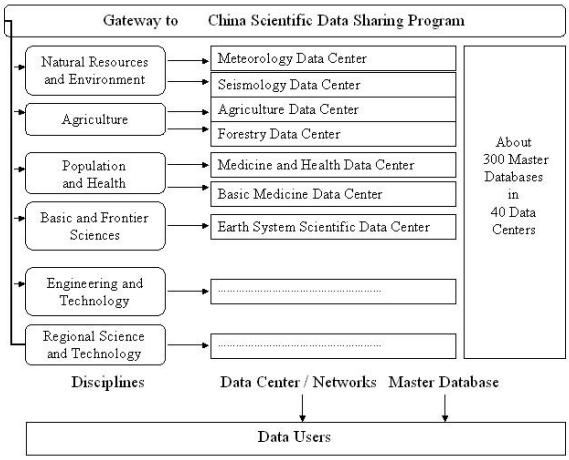

Figure 1. Structure of the Scientific Data Sharing Program

China's progress in scientific data sharing can be tracked through the development of its policy and legal framework, data standards, pilot projects and international cooperation. Open access to scientific data cannot be achieved without an implementing policy and legal framework. MoST initiated the policy-making process at the beginning of the SDSP. By the end of 2006, there were four laws and regulations being drafted at the national level, and 39 rules and regulations that were either being drafted or already released by the relevant departments and agencies.

Because the SDSP involves massive scientific data resources, it is difficult to reach the goal of effective sharing without unified data and technology standards. Of the 32 principal standards identified under the Program, 23 have been completed. This has been done based mainly on the analysis of standards at the national and international levels. Training courses on the implementation of these standards have been conducted. Based on these principal standards, the relevant government agencies have established more than 120 data management standards for different sectors.

The implementation of the SDSP goals has been facilitated as a result of these standard-setting activities. Twenty-four government agencies have been engaged in the Program so far. Despite the differences in scale and status of progress, the scientific data centers or networks that have been established as pilot projects have an effective structure and cover a wide range of topics. They include data centers in Meteorology, Surveying and Mapping, Hydrology and Water Resources, Seismology, Oceanography, Land and Natural Resources, Agriculture, and Forestry, as well as networks for Medicine and Health Data Sharing, Earth System Science Data Sharing, and Sustainable Development Information.

International cooperation plays an important role in the development of data sharing policies, technology and data sources of China. In June of 2004, a workshop on Strategies for Open Access to and Preservation of Scientific Data was held in Beijing, with over 100 participants from 12 countries. Several study tours to the United States and the European countries were organized in the 2002-2005 period for investigation and exchange of views on the management of scientific data resources. Scientific data centers and networks in China also establish links with other countries and regions for data sharing and bring in large amounts of data resources from around the world. During the 20th CODATA International Conference in 2006, MoST co-organized the key session on Global Scientific Data Sharing and Application, promoting the understanding of open access to scientific data among the world's scientific data community.

3 RESULTS OF THE SCIENTIFIC DATA SHARING PROGRAM

After nearly five years of efforts, the concept of "scientific data sharing" has become more widely accepted by the Chinese scientific community, and gradually more popular among government departments, institutions and the general public. For instance, about 180 thousand hits were found on Google with the Chinese characters for "scientific data sharing" by the end of August, 2006, and that number continues to increase.

The data standards are starting to have an impact as well. The 23 principal standards form a foundation for the standardization of scientific data sharing to develop gradually in a top-down approach. Training courses on the principal standards also promote data sharing in various sectors. The principal standards have created the necessary conditions for high-volume data sharing and high-speed network connections between the SDSP projects and other application systems.

The data integration and sharing activities have provided greater value to the more than 25 billion yuan of data

resources produced by the government. By December 2005, SDSP pilot projects had integrated and upgraded 864 databases with nearly 50TB data, which comprise about one-third of the existing public-welfare and basic scientific data in China. For example, in the 1980s and 1990s, there were nearly 100 domestic databases of Traditional Chinese Medical Science and Medicine, but the total amount of data was less than 3GB. Since 2002, most of these databases have been integrated into eight and the amount of data has increased to 32GB.

As a result of these efforts in the pilot stage of the Program, the effects of scientific data sharing are evident. By the end of 2005, there were over 50 thousand registered users, of which 14 million were accessing users, and about 15TB of data were downloaded. Analysis by MoST shows that over half of the users are from universities and scientific research institutes, and the data obtained are used mainly in scientific research, education, science popularization and the formulation of various plans. Over 1225 national research key projects have benefited from the Program.

These initial results have attracted the attention of international data organizations. In July 2005, the World Data Center system evaluated nine associated scientific data centers in China, which concluded in a positive assessment of their progress. After five years of training, the government departments and agencies involved have formed professional teams with people specialized in data standards, data sharing policy, database rebuilding, data services, data analysis, data websites and other related functions. In addition, a great number of graduate students are being trained.

4 FUTURE CONSIDERATIONS

Although prominent achievements have been obtained in recent years on scientific data sharing in China, some problems still exist in the following three aspects.

First, relevant laws and regulations need to be established, or improved and put into practice. Policies are insufficient or need to be perfected for encouraging data sharing, effective long-term operations and evaluation. Second, the gaps between the SDSP's principal standards and standards in specific disciplines need to be filled. Some domains lack systematic data sharing standards. And third, authentic data resources need to be further integrated. Effective integration of dispersed data resources is necessary for building world-class databases.

To solve the above-mentioned problems, high priority should be given to the following action areas.

The national policy and legislation system is indispensable for promoting open access to scientific data. Policies and laws should regulate the procedures of data collection, integration, sharing and utilization; harmonize the interests between data owners, management staff and users; and safeguard the sound and sustainable development of data sharing.

Scientific data that originate from scientific activities of various types have different formats and complexities. Therefore, a uniform, standard system must be established to guide data integration and exchange services, and to improve the quality of data resources.

According to the SDSP principles, it is crucially important to build national scientific data sharing centers. MoST plans to expand and establish a number of national scientific data centers and networks, and the national databases will be generated from these centers. The construction of data resources will focus on integrating data from government produced, owned and funded projects, in order to facilitate the effective and wide uses of the data by the whole society.

China is committed to the policy of reform and opening up, and to learning from the advanced experiences in other countries. Our country wishes to introduce mature mechanisms, standards and criteria of foreign data management. It encourages Chinese scientists to go abroad, to strengthen and broaden cooperation with international scientific organizations and the science and technology communities of all countries, and to join hands in developing a grand international platform for data sharing in this information age. There are many potential avenues for future international cooperation, such as to carry out exchanges, research and discussions on policies, with an emphasis on the evaluation of achievements, incentives and effective long-term operational mechanisms; to encourage bilateral and multilateral projects, with a focus on international exchange and sharing of data in various domains; and to promote the training of talented people and technology exchange on policy research, standards, database technology, data analysis and network services.

5 CONCLUSION

As economic globalization and the international scientific and technical activities increase constantly, the need for the exchange and comprehensive uses of scientific data has grown ever stronger. Data sharing will become an effective way to promote science and innovation, and thus become an inevitable choice for progress in the information age.