Joint Study on Improving Scientific Data Sharing Between the United States and China

U.S. National Committee for CODATA
Board on Research Data and Information
The National Academies, U.S.
and
Chinese National Committee for CODATA
International Cooperation Bureau
Chinese Academy of Sciences

STUDY PROPOSAL

SUMMARY
The National Research Council's Board on Research Data and Information (BRDI), which also serves as the U.S. National Committee for CODATA (USNC/CODATA), proposes to conduct a study of the various kinds of barriers to cooperation in scientific data activities between the institutions and individual researchers in the United States and the People's Republic of China (the study), and to make recommendations to help eliminate or minimize the impact of such barriers. The study will be conducted jointly by USNC/CODATA and the Chinese National Committee for CODATA, under the auspices of their respective Academies of Sciences. The study will be performed over a period of 18 months, beginning August 1, 2010, pursuant to the following statement of task:

1. Identify areas of joint research in the earth and environmental, and the health and biomedical areas, that are particularly data intensive and would benefit from greater exchange and sharing of data.
2. Analyze the barriers to data sharing or exchange, including those that are scientific and technical; institutional; economic and financial; legal or policy-based; and socio-cultural. Identify barriers that are based on legitimate countervailing concerns and those that are candidates for reduction or elimination.
3. Produce a consensus report with conclusions and recommendations for consideration by both countries to overcome excessive or unnecessary barriers and to improve data sharing, both nationally and internationally.

A joint report will be published in both English and Chinese and actively disseminated within each country's relevant research and policy communities. Both editions will be copyrighted by the NAS.

Intellectual Merit of the Proposed Activity

Rapidly changing technological capabilities for creating, manipulating, disseminating, and using digital scientific data are producing many new opportunities and challenges. The opportunities arise primarily in data-intensive research and applications, in the integration of heterogeneous data for new results, and in making vast amounts of factual information available for a broad spectrum of users for collaborative research. The inherent challenges are in effectively managing

these data resources for optimal access and use, and for developing rational rules and structures for such processes.

Both the United States and China have vigorous ongoing and planned scientific data collection and related research activities, using data that are collected operationally by government agencies or through research either within the government or in academia with government funding. In recent years, both the Chinese and U.S. governments have rapidly modernized their publicly funded research systems and produced a great deal of scientific data. These data collections are also potentially valuable for cooperative studies of global health, earth and space sciences, geospatial data applications, and various types of basic and applied research generally. Improved access to the Chinese data resources and expertise is in the interest of the U.S. research and applications communities, and vice versa. The proposed study will promote these interests by jointly studying the various scientific and technical; institutional and management; economic and financial; legal and policy; and normative and socio-cultural barriers to data sharing, and to bilateral cooperation in scientific data and information, and by making recommendations for removing or reducing such barriers to the relevant institutional and governmental managers in each country.

BACKGROUND

Between 2000 and 2005, the U.S. National Committee for CODATA and the Chinese National Committee for CODATA held a series of bilateral meetings with senior science officials and data managers from both countries to discuss various data management and policy issues. Particularly noteworthy in this regard was the high-level Scientific Data Sharing Program announced in February 2003 by the Chinese Ministry of Science and Technology (MOST), and supported by the National People's Congress. This science policy initiative was substantially shaped by and informally attributed to the results of the bilateral CODATA meetings. Also important in this regard was the more recent liberalization of the public sector information law with the enactment of national freedom of information act in 2008.

The two CODATA Committees held a bilateral workshop in 2004 in Beijing, focused on the preservation and sharing of scientific data and information. This meeting re-confirmed the commitment of the Chinese science policy community to promoting greater openness regarding Chinese research data (NRC, 2006).

In 2006, the two CODATA Committees, under the auspices of their respective Academies of Sciences, established the U.S. - China Roundtable on Scientific Data Cooperation (the Roundtable), to convene a series of meetings over an initial three-year period pursuant to the following statement of task:

1. Provide a unique bilateral forum for government, academic, and private-sector stakeholders in the United States and China to discuss and address scientific data practices and policies, pursuant to a mutually agreed agenda.
2. Serve as a catalyst and coordinating body for bilateral cooperation on scientific data practices and policies at the Academy and national level in each country, with appropriate recognition and representation of other thematically related bilateral and international activities.

The Roundtable participants identified four thematic areas of mutual interest in which there could be joint bilateral projects on data management and sharing. The four areas identified were: a) health and biomedical data, b) environmental and geospatial data, c) advanced cyber-infrastructure data applications, and d) scientific data policy.

Three Roundtable meetings have been convened in 2006, 2007, and 2009 (in Beijing, Washington, DC, and Qingdao, respectively) to discuss and propose joint projects in these thematic areas. The participants in these Roundtable meetings from each country included researchers, data managers, and decision makers from various government agencies, particularly those who have been either actively involved in bilateral projects in China or would like to initiate joint bilateral projects.

In the first two Roundtable meetings held in Beijing and Washington DC, the participants discussed and identified various technical, management, socio-cultural, and policy of barriers to data sharing. These barriers provided a major challenge to initiation and successful completion of various bilateral projects proposed by scientists on both sides. At the end of the most recent Roundtable meeting held in Qingdao in March 2009, the U.S. as well as the Chinese participants affirmed that there is a compelling need to have data sharing activities in support of scientific cooperation between U.S. and the Chinese scientists. They also recognized that this cooperation is being hampered by various barriers and that there is an urgent need to address them. A joint, bilateral study is therefore proposed that would make recommendations to help remove these barriers, or at least reduce their effect on scientific cooperation across the types of research fields that have been the focus of the bilateral Roundtable. Such a study will build on the experiences and network of experts engaged or affiliated with the Roundtable and accentuate the need for improved data sharing in joint cooperation to research policy-makers and to the scientific communities in each country in addressing high-priority research questions. A list of National Research Council references is included in the Bibliography at the end of the proposal.

The proposed study will identify and analyze barriers to data cooperation using data that are collected operationally by government agencies or through research either within the government or in academia with government funding, from the perspective of several areas, as summarized briefly below.

1. Scientific and technical aspects. Sharing of data and information requires accommodating the needs and practices of different scientific disciplines, as well as encouraging the development of interdisciplinary research values and methods. There likely will be differences among the mandates and objectives of individual institutions for different types of data (e.g., observational vs. experimental, physical science vs. biological science, human subjects or not) and may have disparate procedures and metrics for data quality, and differing criteria for selecting data for sharing with others. The development of databases also depends largely on discipline-independent technology and infrastructure requirements appropriate to the different goals of access and sharing. These include, among others: the development and adoption of common metadata standards and practices; flexible search and retrieval capabilities; technological and semantic interoperability; and appropriately accommodating the evolution in technology (hardware and software), as well as data and information collected in proprietary formats and commercial databases.

2. Institutional aspects. Maintaining data and information as community resources for purposes of sharing and joint scientific activities requires the implementation of effective operational procedures and practices. The institutions need to manage the collection, storage and sharing of data effectively. Other operational issues may include properly managing the volume of data, which is enormous and growing, even in the developing world; coping with the diversity of sources, formats, and documentation; and maintaining a sufficiently long time horizon for access in the face of continually changing definitions, digital media and formats, and hardware and software obsolescence. Planning and developing requirements for data management and sharing must accommodate the continual change and evolution in the practice of science; the local variability in focus, practice, and available technology and other physical and human infrastructure; and the differing mandates and objectives of various data producing institutions, as well as a diversity of potential sharing partners, including scientists, educational institutions, and policy-makers within and outside each country.

3. Economic and financial aspects. Another major barrier to the establishment of a well-managed data center geared towards data sharing is adequate funding. Data centers are not among the most pressing recognized priorities in either country, despite the importance, and considerable potential contributions, of well-managed scientific information resources to research capacity building and to social and economic development. Creative and well-planned approaches could reduce the financial burden. Moreover, the potential social and economic returns from sharing of data that is relevant to high-priority national and global problems that need collaborative solutions can more than offset the relatively small financial costs of data sharing and exchanges.

4. Legal and policy aspects. Most research databases and data centers in China are managed directly by government ministries and subject to a relatively restrictive state information regime, whereas the situation in the United States is much more decentralized and diffuse, coupled with generally much more open access policies and laws. Laws and policies can provide a major challenge or opportunity for data sharing. They tend to be founded on deeply rooted political, institutional, and cultural factors, some of which apply to the overall public information regime and some of which are exacerbated in the scientific context by perceived political or economic sensitivities of the subject matter (e.g., domestic disease statistics, biodiversity information, environmental degradation or resource exploitation, or the disclosure of many otherwise personally or nationally sensitive facts).

5. Socio-cultural aspects. Various underlying scientific and national norms and socio-cultural values can have important effects on data sharing behavior similar to the legal and policy influences. At the scientific community level, these factors include traditional data sharing behavior in each scientific sub-community, the adequacy of recognition of the importance of the scientific activities by the scientists and their institutions, and the existence or absence of financial incentives or rewards. Relevant cultural and normative values also exist at the broader societal level in each county that need to be understood.

PLANNED ACTIVITIES

This study will be conducted pursuant to the following Statement of Task:

1. Identify areas of joint research in the earth and environmental, and the health and biomedical areas, that are particularly data intensive and would benefit from greater exchange and sharing of data.
2. Analyze the barriers to data sharing or exchange, including those that are scientific and technical; institutional; economic and financial; legal or policy-based; and socio-cultural. Identify barriers that are based on legitimate countervailing concerns and those that are candidates for reduction or elimination.
3. Produce a consensus report with conclusions and recommendations for consideration by both countries to overcome excessive or unnecessary barriers and to improve data sharing, both nationally and internationally.

The study will be conducted jointly by the U.S. committee appointed by the NRC Chair and a separate Chinese committee appointed by the Chinese Academy of Sciences, in accordance with the procedures of the respective institutions. The committees will be composed of an approximately equal number of members from each nation, with relevant expertise, work experience, and range of perspectives on the issues to be addressed in the joint study. Such expertise will include the scientific, technical, institutional, management, financial, legal, policy, and sociological factors associated with scientific data cooperation between the United States and China. Nominees to the NRC committee will be particularly sought from the membership of the NAS, NAE, and IOM.

Together, the two committees will oversee all the study activities, including the conduct of the workshops, the meetings, the research activities, and the drafting of the study report. The study report will be drafted in English and reviewed jointly by U.S. and Chinese reviewers under the NRC Report Review Committee rules and auspices.

There will be two workshops and three writing sessions as part of the study. One workshop will be held in the United States and one in China, with each workshop followed by a meeting of the joint study committees to draft the report. A third drafting meeting will be held after the first two workshops to complete the writing of the report. A symposium to convey the results of the completed study will be held in Beijing once the report is released in the United States.

In conducting the study, a questionnaire will be used as noted below to gather information for the committees on possible barriers to data sharing in the discipline areas covered by the study. Additional research will be performed by the staff between the meetings at committee direction. A research consultant will draft a background paper prior to the first meeting of the study committees, addressing the statement of task and identifying and analyzing the different barriers to data sharing.

Study Committees joint meeting and public Workshop #1: The first workshop will discuss the various scientific and national benefits of data sharing and identify the types of barriers experienced by researchers and scientific institutions in the U.S. and China. There will be a number of invited presentations on different aspects of data sharing and background issues by experts with experience in bilateral scientific projects or programs between the United States and China. The sponsors of the study also will be invited to present their experiences and activities.

The workshop will be followed by a joint committee meeting to discuss information gathered at the workshop and develop a detailed outline of the report. Finally, the committees will select some past, current, or failed bilateral research and data sharing projects that could be used as case studies to highlight the nature of the barriers and help to suggest solutions to overcome these barriers. The chapters of the report will be drafted by small teams of Chinese and U.S. committee members.

Study Committees joint meeting and public Workshop #2: The second workshop will further explore the barriers to data sharing in the categories mentioned earlier, distinguishing between barriers that are common to bilateral projects across all disciplines and those that arise only in a particular discipline or research context. A meeting of the joint study committees will follow the workshop to synthesize the information from the workshop, examine the results of the case studies, review the draft text of the report, discuss preliminary conclusions and recommendations, and identify any issues that still need to be addressed.

Study Committees joint meeting #3: The study committees will meet a third time to agree on the conclusions and recommendations, and complete the report text and any outstanding issues. The final text and the review draft will be completed by email exchanges or the use of a password-protected wiki.

Symposium: After the report is released, a one-day public symposium will be held in China discuss the study's conclusions and recommendations with high-level invited policy and research representatives. A smaller sponsor briefing and release event will be held in Washington just prior to the Beijing symposium to release the report in the United States.

The work of the NRC committee will be subject to the institutional procedures implementing the requirements of FACA Section 15. In this context, members of the Chinese committee and staff will be designated by the NRC as consultants so that they can participate in closed meetings with the NRC committee and staff.

Outreach and Communication

The study workshops, the final meeting, and of course the report of the study itself will constitute a major effort toward outreach and communication about these barriers to bilateral scientific cooperation within the research establishment and government officials of the two countries. A major effort will be made to identify the decision makers and policy makers at the institutional and government levels and to invite them to attend the workshops and the final meeting, as well as to subsequently follow up on implementing the study recommendations. The joint study committee co-chairs, members, and the project director also will be encouraged to give presentations subsequently about the results of the study at various professional meetings and conferences, including the intergovernmental working groups, the Joint Bilateral Commission under the Science and Technology Agreement between the United States and China, and events related to bilateral scientific cooperation.

Substantial efforts will be made to communicate information about the workshops, meetings, the questionnaire, and the results of the study via the websites of the collaborating organizations and

the sponsors. Various media outlets also will be targeted in both countries.

Study Report Dissemination
The study report will be available to the public and widely disseminated, without restriction, including publication on the National Academies Press Web site (www.nap.edu). Print copies of the study report will be prepared in sufficient quantity to ensure its distribution to the sponsors and other relevant parties, in accordance with Academy policy. It is also expected that the Chinese collaborators will publish and disseminate the study report in the Mandarin language.

Collaborations with Other Organizations
Bilateral understanding of the barriers to data sharing and joint actions that need to be taken to resolve them are major reasons why this study is being planned. As has already been noted, the study is a collaborative effort of the two national CODATA committees and their Academies. In addition, many governmental agencies and academic institutions in the U.S. have gained extensive knowledge and experience in bilateral projects in China and will be valuable resources for relevant information and expert participants. Some U.S. agencies, such as the NSF and the NIH, have offices in China to assist in bilateral scientific cooperation. These specialists will be valuable resources for guidance on all aspects of the study, including identification of the issues, workshop speakers and attendees. Other bilateral research projects and organizations will be consulted as well.

The study will be conducted with the participation of many U.S. and Chinese governmental agencies, research institutions managed by the Chinese Academy of Sciences (CAS), and other academic institutions and universities. Most of the organizations that will be asked to provide information to the study have already participated in the three U. S. - China Roundtable meetings held so far, and have indicated an interest in overcoming the barriers to data sharing cooperation.

Public Information about the Project
In order to afford the public greater knowledge of Academy activities and an opportunity to provide comments on those activities, the Academy may post on its website (http://www.nas.edu) the following information as appropriate under its procedures: (1) notices of meetings open to the public; (2) brief descriptions of projects; (3) committee appointments, if any (including biographies of committee members); (4) report information; and (5) any other pertinent information.

Responsible Staff in the United States
Subhash Kuvelker, J.D.
Senior Program Officer, Board on Research Data and Information
and the U.S. National Committee for CODATA
National Academy of Sciences
Washington, DC
puhlir@nas.edu

Responsible Staff in China
LI Jianhui
Director, Scientific Data Base

Computer Network and Information Center
Chinese Academy of Sciences
Beijing
lijh@sdb.cnic.cn

Bibliography of National Research Council Reports

The Socioeconomic Effects of Public Sector Information on Digital Networks (2009).

Ensuring the Integrity, Availability, and Stewardship of Research Data in the Digital Age (COSEPUP, 2009).

Environmental Data Management at NOAA: Archiving, Stewardship, and Access (2007).

Strategies for Preservation of and Open Access to Scientific Data in China (2006).

Open Access and the Public Domain in Digital Data and Information for Science (2004).

Licensing Geographic Data and Services (2004).

The Role of Scientific and Technical Data and Information in the Public Domain (2003).

Sharing Publication-related Data and Materials: Responsibilities of Authorship in the Life Sciences (2003).

Access to Research Data in the 21st Century (2002).

Health Data in the Information Age: Use, Disclosure, and Privacy (2002).

Improving Access to and Confidentiality of Research Data (2000).

A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases (1999).

Bits of Power: Issues in Global Access to Scientific Data (1997).

Sharing Research Data (1985).