Workshop on Big Data

for International Scientific Programmes: Challenges and Opportunities

8-9 June 2014 Beijing China

PROGRAMME BOOK

Sponsor

Co-Sponsors

Organizer

CONTENTS

I. Sponsoring Organizations

International Council for Science (ICSU)

The International Council for Science (ICSU) is a non-governmental organization with a global membership of national scientific bodies (121members, representing 141 countries) and international scientific unions (31 members). ICSU mobilizes the knowledge and resources of the international scientific community to strengthen international science for the benefit of society.

ICSU: *www.icsu.org*

Committee on Data for Science and Technology (CODATA)

CODATA, the ICSU Committee on Data for Science and Technology, was established in 1966 to meet a need for an international coordinating body to improve the management and preservation of scientific data. CODATA has been at the forefront of data science and data policy issues since that date.

CODATA supports ICSU's mission of 'strengthening international science for the benefit of society' by 'promoting improved scientific and technical data management and use'. CODATA achieves this mission through three strands of activity:

1. Leading an international policy agenda for open scientific data (through the CODATA International Data Policy Committee and partnerships with international programmes);
2. Coordinating work to confront emerging challenges and 'hot topics' at the frontiers of data science (through CODATA Task Groups and Working Groups and other initiatives).
3. Developing data strategies for international science programmes and supporting ICSU activities such as Future Earth and Integrated Research on Disaster Risk (IRDR) to address data management needs.

Through events like the Workshop on Big Data for International Scientific Programmes and SciDataCon 2014, CODATA collaborates with international partners to break new ground in addressing the data issues at the heart of key research questions.

CODATA: *www.codata.org*
SciDataCon 2014: *www.scidatacon2014.org*

ICSU World Data System (WDS)

The World Data System (WDS) builds on 50-year legacy of the World Data Centres and Federation of Astronomical and Geophysical data analysis Services established by the International Council of Science (ICSU) to manage data generated by the International Geophysical Year (1957–58). These bodies were later disbanded by the ICSU General Assembly in 2008 and replaced by WDS in 2009.

WDS aims at facilitating the scientific research endeavours by coordinating trusted scientific data services for the provision, use, and preservation of relevant datasets. To fulfil its remit, WDS is striving to build worldwide 'communities of excellence' for scientific data services by certifying member organizations—holders and providers of data or data products—from wide-ranging fields using internationally recognized standards. WDS Members are then the building blocks of a searchable common infrastructure with which to form a data system that is both interoperable and distributed.

WDS: *www.icsu-wds.org*

Future Earth

Future Earth is the global research platform providing the knowledge and support to accelerate our transformations to a sustainable world. Launched in June 2012 at the UN Conference on Sustainable Development (Rio+20), Future Earth is a 10-year international programme that will provide critical knowledge required for societies to face the challenges posed by global environmental change and to identify opportunities for a transition to global sustainability. Future Earth integrates research activities under existing international Global Environmental Change programmes (DIVERSITAS, the International Geosphere-Biosphere Programme (IGBP), the International Human Dimensions Programme (IHDP) and the World Climate Research Programme (WCRP)). Future Earth is an international hub to coordinate new, trans-disciplinary approaches to research, integrating different disciplines from the natural and social sciences (including economic, legal and behavioural research), engineering and humanities. It is also a platform for international engagement to ensure that knowledge is generated in partnership with society and users of science. Future Earth is sponsored by the members of the Science and Technology Alliance for Global Sustainability comprising the International Council for Science (ICSU), the International Social Science Council (ISSC), the Belmont Forum of funding agencies, the United Nations Educational, Scientific, and Cultural Organization (UNESCO), the United Nations Environment Programme (UNEP), the United Nations University (UNU), and the World Meteorological Organization (WMO) as an observer.

Future Earth: *www.futureearth.info*

Workshop on Big Data for International Scientific Programmes

Integrated Research on Disaster Risk (IRDR)

The Integrated Research on Disaster Risk (IRDR) programme is a decade-long integrated research initiative co-sponsored by the International Council for Science (ICSU), the International Social Science Council (ISSC), and the United Nations International Strategy for Disaster Reduction (UNISDR). It is a global, trans-disciplinary research programme created to address the major challenges of natural and human-induced environmental hazards. The complexity of the task is such that it requires the full integration of research expertise from the natural, socio-economic, health and engineering sciences as well as policy-making, coupled with an understanding of the role of communications, and public and political responses to reduce the risk.

IRDR is governed by a 15-member Scientific Committee (SC) set up by and on behalf of the Co-Sponsors. Its responsibilities are to define, develop and prioritize plans for the IRDR, guide its programming, budgeting and implementation, establish a mechanism for oversight of programme activities, and disseminate and publicize its results. The execution of IRDR programme promotion, coordination and related functions is undertaken by the IRDR International Programme Office (IPO). The IPO is located in Beijing, China and is hosted by the Institute of Remote Sensing and Digital Earth (RADI) of the Chinese Academy of Sciences (CAS). Operational funds are provided by the Chinese Association of Science and Technology (CAST).

IRDR: *www.irdrinternational.org*

Research Data Alliance (RDA)

The Research Data Alliance (RDA, rd-alliance.org) was launched in March 2013 with the support of the European Commission, the US National Science Foundation and the Australian Government through the Australian National Data Service (ANDS) as an international, community-powered organization. RDA's vision is a world where researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society. RDA's mission is to build the social and technical bridges that enable data sharing. This is being accomplished through the creation, adoption and use of the social, organizational, and technical infrastructure designed to reduce barriers to data sharing and exchange with Working and Interest groups as the main RDA global co-operation mechanism. Membership is free and open to all on *www.rd-alliance.org*.

Group on Earth Observations (GEO)

The Group on Earth Observations (GEO) is a voluntary partnership of governments and organizations that envisions "a future wherein decisions and actions for the benefit of humankind are informed by coordinated, comprehensive and sustained Earth observations and information." GEO Member governments include 89 nations and the European Commission, and 77 Participating Organizations comprised of international bodies with a mandate in Earth observations. Together, the GEO community is creating a Global Earth Observation System of Systems (GEOSS) that will link Earth observation resources world-wide across multiple Societal Benefit Areas, including agriculture, biodiversity, climate, disasters, ecosystems, energy, health, water and weather, and make those resources available for informed decision-making.

GEO: *www.earthobservations.org*

International Society for Digital Earth (ISDE)

The International Society for Digital Earth (ISDE) was founded in May, 2006 in China, on the principles of the 1999 Beijing Declaration on Digital Earth. The ISDE promotes international cooperation in the Digital Earth Vision, and facilitates Digital Earth technologies to play key roles in, *inter alia*, economic and socially-sustainable development, environmental protection, early warning and disaster mitigation, natural resources conservation, education and improvement of the well-being of the society in general. The mission of the ISDE is to provide a framework for understanding evolving society-beneficial technologies, current and newly emerging, and to revise the Digital Earth vision in light of new developments. The 2014 Digital Earth Summit, themed with "Digital Earth for Education for Sustainable Development", will be organized on Nov. 9-11, 2014 in Nagoya, Japan.

The *International Journal of Digital Earth* (IJDE) is the academic journal of the International Society for Digital Earth, and jointly published by the Taylor & Francis Group. The IJDE was launched in March 2008, and accepted for coverage in the Science Citation Index Expanded (SCI-E) in August 2009. IJDE has now become one of the leading journals in the geospatial information field in the world.

The ISDE Secretariat and the IJDE Editorial Office are hosted by the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences.

ISDE: *www.digitalearth-isde.org*
IJDE: *mc.manuscriptcentral.com/ijde*
2014 Digital Earth Summit: *www.isde-j.com/summit2014*

Workshop on Big Data for International Scientific Programmes

Institute of Remote Sensing and Digital Earth (RADI) Chinese Academy of Sciences (CAS)

The Institute of Remote Sensing and Digital Earth (RADI), Chinese Academy of Sciences (CAS) is a comprehensive research institute committed to the research and applications of leading technologies in Earth observation, geospatial information science and remote sensing information.

Its main objective is to construct and operate spaceborne, airborne and ground-based Earth observation systems that can provide resource-environmental spatial information at the regional and global level, and form a Digital Earth science platform.

RADI operates two national key scientific and technical (S&T) infrastructures: the China Remote Sensing Satellite Ground Station and the Airborne Remote Sensing Aircraft. RADI is home to many cutting-edge research resources, including the State Key Laboratory of Remote Sensing Science, the Center for Applied Technologies of Earth Observation, the National Engineering Center for Geoinformatics, and the CAS Laboratory of Digital Earth Sciences.

RADI is also an active participant in international collaborations, and hosts several international S&T platforms, including the International Centre on Space Technologies for Natural and Cultural Heritage under the auspices of UNESCO, the International Society for Digital Earth, the International Programme Office for Integrated Research on Disaster Risk, and the CAS-TWAS Centre of Excellence on Space Technology for Disaster Mitigation.

RADI: *english.radi.cas.cn*

II. Programme

Workshop on Big Data

for International Scientific Programmes: Challenges and Opportunities
Sunday & Monday, June 8-9, 2014 Beijing, China

To provide a better understanding of the opportunities and challenges of 'Big Data' for international collaborative science programmes, CODATA (the ICSU Committee on Data for Science and Technology) will hold a high-level scientific meeting to discuss the Big Data for science discovery and innovation, and to promote international and cross-disciplinary collaboration, in the age of Big Data.

Sponsor:	The ICSU Committee on Data for Science and Technology (CODATA)
Co-Sponsors:	ICSU World Data System (WDS) Future Earth (FE) Integrated Research on Disaster Risk (IRDR) Research Data Alliance (RDA) Group on Earth Observations (GEO) International Society for Digital Earth (ISDE) Institute of Remote Sensing and Digital Earth (RADI), Chinese Academy of Sciences (CAS)
Organizer:	Institute of Remote Sensing and Digital Earth , CAS
Venue:	Beijing International Convention Center (BICC) *http://www.bicc.com.cn/english/index.html*
Recommended Hotel:	Beijing North Star Continental Grand Hotel *near the Olympic Park*. Address: No. 8 Beichen East Rd., Chaoyang District, Beijing, China *http://yd.bcghotel.com/English/*

Day One: Using Big Data to Advance International Scientific Programmes
Sunday, June 8, 2014

Registration	**08.30-08.50**	**The 3rd Floor of BICC**
Opening Ceremony	**09.00-10.00**	**Chair:** Sara Graves, *Secretary General of CODATA* **Remarks:** • GUO Huadong, *President of CODATA* • GAO Wen, *Vice President of the National Natural Science Foundation of China (NSFC)* • Krishan Lal, *Immediate Past President of the Indian National Science Academy (INSA)* • Barbara Ryan, *Secretariat Director of GEO* • Mark Stafford Smith, *Chair of the Future Earth Scientific Committee* • Milan Konečný, *Vice President of ISDE* • David Johnston, *Chair of IRDR Scientific Committee* • Mustapha Mokrane, *Executive Director of WDS IPO* • XU Guanhua, *Former Minister of the Ministry of Science and Technology of China (MOST)*
Keynotes Session: Setting the Scene, Big Data and International Science	**10.00-11.00**	**Chair:** Sara Graves, *Secretary General of CODATA* **Opening Keynotes:** • **Big Data, Big Science, Towards Big Discovery** GUO Huadong, *President of CODATA* • **Environmental Informatics: Leveraging BIG Data to Solve BIG Problems** Peter Fox, *Chair of the Tetherless World Constellation (TWC) at Rensselaer Polytechnic Institute (RPI)*
Group Photo	**11.00-11.30**	**The 3rd Floor of BICC**
Break		

Keynotes Session: Setting the Scene, Big Data and International Science	**11.30-12.30**	**Chair:** Krishan Lal, *Former President of CODATA* **Opening Keynotes:** • **The Group on Earth Observations (GEO) through 2025** Barbara Ryan, *Secretariat Director of GEO* • **Future Earth – An Agenda Needing Use-Oriented Data** Mark Stafford Smith, *Chair of the Future Earth Scientific Committee*
Lunch	**12.30-13.30**	**The 2nd Floor of BICC**
Session A *Big Data Challenges for International Collaborations*	**13.30-15.00**	**Chair:** John Rumble, *Former President of CODATA* • **A Climate Scientist's Reflection on Big Data** Deliang Chen, *August Chair at the University of Gothenburg; Former Executive Director of ICSU* • **Big Data in Climate Change Adaptation** Jakob Rhyner, *United Nations University Vice Rector in Europe; Director of the Institute for Environment and Human Security at United Nations University* • **Toward Global Cooperation on Microbial Big Data** MA Juncai, *Director of WFCC-MIRCEN World Data Center of Microorganisms (WDCM)* • **Big Data and Big Ideas** Shuichi Iwata, *Professor at the Graduate School of Project Design of Tokyo University; Editor-in-Chief of the CODATA Data Science Journal (DSJ)*
Break	**15.00-15.30**	**The 3rd Floor of BICC**

Session B *Big Data Today: Lessons and Challenges from Science Initiatives and the Commercial World*	**15.30-17.30**	**Chair:** Shuichi Iwata, *Former President of CODATA* • **Genomics: Big Data, Big Science, Big Sharing** YANG Huanming, *President of BGI-China* • **Datafication of the World: People, Community, and Physical Space** Eric Chang, *Senior Director of Technology Strategy at Microsoft Research Asia (MSRA)* • **Chemical, Material, and Physical Property Big Data: Balancing Scientific and Commercial Interests** John Rumble, *President of R&R Data Services in Gaithersburg MD* • **The Observations and Data Revolution: Opportunities for Future Earth** Frans Berkhout, *Interim Director of Future Earth* • **Policy Issues in Big Data** Paul Uhlir, *Director of the Board on Research Data and Information (BRDI) at the U.S. National Academies*
Welcome Banquet	**18.30**	**Xibei Ninty-Nine Yurts**

Day Two: Building International Cooperation and Collaboration on Big Data for International Science
Monday, June 9, 2014

Session C **Invited Talks and Discussion** Big Data Opportunities: How to Make Progress	**09.00-10.30**	**Chair:** Robert Chen, *Former Secretary General of CODATA* • **NIST Big Data Public Working Group & Standardization Activities** Wo Chang, *NIST Data Coordinator; Digital Data Advisor for the NIST Information Technology Laboratory (ITL)* • **Using Big Data Analytics for Knowledge Extraction and Collaboration** Sara Graves, *Secretary General of CODATA*

		• **Big Data Infrastructure and Analysis: How to Make Progress?** Pit Pichappan, *Senior Scientist at the Digital Information Research Foundation (India & U.K.)* • **Data-Intensive Scientific Discovery in Digital Earth** WANG Lizhe, *Professor at RADI, CAS*
Break	**10.30-11.00**	**The 3rd Floor of BICC**
Session D Opportunities for Collaboration on Big Data for International Science	**11.00-12.30**	**Chair:** Paul Uhlir, *U.S. National Committee for CODATA* • **From TerraPop to Teraflops: The Challenge of Integrating Big Social and Natural Science Data** Robert Chen, *Director of the Center for International Earth Science Information Network (CIESIN) at the Earth Institute of Columbia University* • **Data for Research That Connects Population, Agriculture and Environment: Opportunities and Challenges** Myron Gutmann, *Research Professor of the Institute for Social Research at the University of Michigan* • **Big Data: Issues in Organizing and Retrieval** A.R.D. Prasad, *Head of the Documentation Research and Training Centre (DRTC) at the Indian Statistical Institute (ISI)* • **Can We Recommend the Best Data for Researching a Certain Disaster Event? - Multidisciplinary Data Correlation Analysis as Example** LI Guoqing, *Professor at RADI, CAS; Member of WDS Scientific Committee*
Lunch	**12.30-13.30**	**The 2nd Floor of BICC**

Session E Opportunities for Collaboration on Big Data for International Science	**13.30-15.00**	**Chair:** Simon Hodson, *Executive Director of CODATA* • **GigaDB: Open Data for the Big Data Era** Chris Hunter, *Lead Bio-Curator for GigaDB at GigaScience* • **Efficiency of Knowledge Extraction on the Big Data Framework** Jan-Ming Ho, *Research Fellow at the Institute of Information Science, Academia Sinica* • **Force Multiplier for Science through Crowd Source Validation and Visualization** Tim Foresman, *SIBA Chair for Spatial Information at the Queensland University of Technology* • **Global Change Research Data Publishing and Repository in China** LIU Chuang, *Professor at the Institute of Geographic Sciences and Natural Resources Research (IGSNRR), CAS*
Break	**15.00-15.30**	**The 3rd Floor of BICC**
Closing Session	**15.30-17.00**	**Panel Discussion: An Agenda for Big Data and International Science** **Chair:** GUO Huadong, *President of CODATA* Robert Chen, *Former Secretary General of CODATA* **Short statements. Panel to comprise:** • Sara Graves, *Secretary General of CODATA* • Peter Fox, *Chair of Tetherless World Constellation (TWC) at Rensselaer Polytechnic Institute (RPI)* • Barbara Ryan, *Secretariat Director of GEO* • YANG Huanming, *President of BGI-China* • Wo Chang, *Co-chair of RDA Big Data Infrastructure Working Group* • A.R.D. Prasad, *Head of the Documentation Research and Training Centre (DRTC) at the Indian Statistical Institute (ISI)* • LI Guoqing, *Member of WDS Scientific Committee* **Closing Remarks:** GUO Huadong, *President of CODATA*

III. Remarks and Abstracts

Opening Remarks:

GUO Huadong
President of CODATA

GAO Wen
Vice President of the National Natural Science Foundation of China (NSFC)

Krishan Lal
Immediate Past President of the Indian National Science Academy (INSA)

Barbara Ryan
Secretariat Director of GEO

Mark Stafford Smith
Chair of the Future Earth Scientific Committee

Milan Konečný
Vice President of ISDE

David Johnston
Chair of IRDR Scientific Committee

Mustapha Mokrane
Executive Director of WDS IPO

XU Guanhua
Former Minister of the Ministry of Science and Technology of China (MOST)

Presentation Abstracts:

Keynotes Session - Setting the Scene, Big Data and International Science

Big Data, Big Science, Towards Big Discovery

GUO Huadong
President of CODATA

Big Data encompasses datasets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process within a tolerable elapsed time. It is more than just a problem of space—Big Data has been recognized as spawning one of the most daunting challenges facing modern science: how to cope with the flood of data now being generated. Given that data-intensive science has now been widely accepted as “the fourth paradigm” of science, it is important to discuss how to understand Scientific Big Data for knowledge discovery.

The most highlighted fields of Scientific Big Data for knowledge discovery is big science. Big science is characterized by large-scale instruments and facilities, funding from government or international agencies, and research conducted by teams or groups of scientists and technicians. Some of the best-known big science projects include the Large Hadron Collider experiments, the Sloan Digital Sky Survey, the Human Genome Project and the Global Change Project. Modern big science in turn features big datasets. For example, continuing at a rate of about 200 GB per night, SDSS has amassed more than 140 terabytes of information. The Large Synoptic Survey Telescope, successor to SDSS, will come online in 2016 and is anticipated to acquire that amount of data every five days. The data flow in LHC would exceed 150 million petabytes annually, or nearly 500 exabytes per day, before replication.

We have already had a number of great scientific discoveries through big science. It is highly expected that more exciting scientific discoveries will be found with the aid of Scientific Big Data.

Environmental Informatics: Leveraging BIG Data to Solve BIG Problems

Peter Fox
Chair of the Tetherless World Constellation (TWC)
at Rensselaer Polytechnic Institute (RPI)

It is becoming increasingly important for scientists to develop a deeper understanding around scientific challenges that affect our society, our economy, and our health. Current environmental questions are those such as how marine ecosystems function, what are the regional and global effects of global change, what is Earth's carbon budget/ledger, etc.? What has become clear is that progress will require scientists and other key stakeholders apply a systems approach to understand these complex, inter-disciplinary problems.

But the barriers are daunting. One problem is the increasingly large amount of data that need to be integrated but the second, more so, is how to deal with the enormous heterogeneity of these data and accompanying application tools and the walls of language, jargon and idiosyncracies that have been unavoidably built up around the understanding of these data by scientists and other consumers working in different scientific fields in isolation for decades.

This talk with present how escience and informatics are changing the fundamental nature of how environmental and social problems of 2020 may be addressed. To overcome these barriers requires new approaches; frameworks, methodologies, and tools. The integration of these datasets must generate data and information products that are meaningful to a variety of researchers as well. The science required to develop such tools is the science of informatics.

The Group on Earth Observations (GEO) through 2025

Barbara Ryan
Secretariat Director of GEO

Ministers from the Group on Earth Observations (GEO) Member governments, meeting in Geneva, Switzerland in January 2014, unanimously renewed the mandate of GEO through 2025. Through a Ministerial Declaration, they reconfirmed that GEO's guiding principles of collaboration in leveraging national, regional and global investments and in developing and coordinating strategies to achieve full and open access to Earth observations data and information in order to support timely and knowledge-based decision-making – are catalysts for improving the quality of life of people around the world, advancing global sustainability, and preserving the planet and its biodiversity.

Five key areas of activities for the next decade include the following: 1.) Advocating for the value of Earth observations and the need to continue improving Earth observation worldwide; 2.) Urging the adoption and implementation of data sharing principles globally; 3.) Advancing the development of the GEOSS information system for the benefit of users; 4.) Developing a comprehensive interdisciplinary knowledge base defining and documenting observations needed for all disciplines and facilitate availability and accessibility of these observations to user communities; and 5.) Cultivating global initiatives tailored to meet specific user needs. The work in these five areas will build on the current Global Earth Observation System of Systems (GEOSS) achievements and ensure that these achievements are both sustained and evolve in keeping pace with policy, technological and information changes at the global level.

Certainly much has been accomplished in GEO's first decade. Yet, more remains to be done. Many - possibly most - nations are facing challenges in operating and sustaining, not to mention expanding, their Earth observation networks. Broad, open data-sharing policies and practices are still not universally accepted and employed. And, communicating scientific results so that policy makers and the general public can understand the long=term (as well as short-term) impacts and implications remains challenging. GEO Members and Participating Organizations must continue to work aggressively to address each of these challenges if Earth system science is going to fully address the significant environmental issues facing the world today.

Future Earth – An Agenda Needing Use-Oriented Data

Mark Stafford Smith
Chair of the Future Earth Scientific Committee

Future Earth is a 10-year international research programme providing the knowledge and support to accelerate our transformations to a sustainable world. Future Earth integrates research activities from the existing international Global Environmental Change programmes (DIVERSITAS, the International Geosphere-Biosphere Programme (IGBP), the International Human Dimensions Programme (IHDP), as well as linking to activities under the World Climate Research Programme (WCRP)). Future Earth is becoming an international hub to coordinate new, trans-disciplinary approaches to research, integrating different disciplines from the natural and social sciences (including economic, legal and behavioural research), engineering and humanities. It will also be a platform for international engagement to ensure that knowledge is generated in partnership with society and users of science. Future Earth is sponsored by the members of the Science and Technology Alliance for Global Sustainability comprising the International Council for Science (ICSU), the International Social Science Council (ISSC), the Belmont Forum of funding agencies, the United Nations Educational, Scientific, and Cultural Organization (UNESCO), the United Nations Environment Programme (UNEP), the United Nations University (UNU), and the World Meteorological Organization (WMO) as an observer (for more information on Future Earth, see *www.futureearth.info*). This talk will introduce Future Earth a little and discuss the importance of linking to (and not duplicating) use-oriented data activities in support of this solutions-oriented agenda.

Session A - Big Data Challenges for International Collaborations

A Climate Scientist's Reflection on Big Data

Deliang Chen
August Chair at the University of Gothenburg; Former Executive Director of ICSU

Science needs data, big and small. For weather and climate research Big Data (referring to the three dimensions that Big Data spans: volume, velocity and variety) has been part of the life for long time, relative to the technology available. How successful is the weather and climate community in dealing with Big Data? What are the lessons learnt and to be learnt? In this presentation I will walk you through the history of data related activity in weather and climate research with a focus on international research collaborations and reflect on the critical success factors for these data related activities. These factors include shared and well-defined missions, persistent efforts, strong national interest and dependence on each other, well organised international coordination through the World Meteorological Office (WMO) on data collection, management and exchange, tight links between research and operation, the complementary role of international research and data programs, appropriate technical and infrastructure support, successful outreach efforts, close cooperation between private and public agencies, and a tradition of constantly challenging technical limits.

Big Data in Climate Change Adaptation

Jakob Rhyner
United Nations University Vice Rector in Europe; Director of the Institute for Environment and Human Security at United Nations University

Data are the basis of any evidence-based planning and decision making. The particular interest in Big Data arises from different angles, particularly from (i) the increasing possibilities to acquire them (e.g. by Earth system observation) or generate them (e.g. by models), but also the use them (e.g. by Earth system models). Climate observation and modeling and climate change adaptation are areas where there is a particular need in Big Data, particularly when it comes to worldwide assessments. The talk will discuss several aspects on data acquisition and generation, their availability, quality, and use. A particular emphasis will be placed on recent Big Data initiatives in the UN system.

Toward Global Cooperation on Microbial Big Data

MA Juncai
Director of WFCC-MIRCEN World Data Center of Microorganisms (WDCM)

Microbial resources are one of the most important natural resources in the world, which provides the scientific basis to support the development of biotechnology and life sciences. Culture collections are like the libraries of these living microbial materials. A hub which can combine these culture collections in different countries has become a necessity. The World Federation for Culture Collection (WFCC) is the community of those long-term conservation and research facilities that brings together more than 600 collections in 68 countries. The WFCC-MIRCEN World Data Centre of Microorganisms (WDCM), the heart of WFCC, once hosted in Australia and Japan, plays as a hub providing a database of microorganisms, analysis of its function and a platform of communication.

WDCM has developed an online reference strain catalogue which helps users to find local sources of the reference strains by citing all collections and providing contact details and the collection's unique reference. Furthermore, WDCM is developing an Analyzer of Bio-Resources (ABC) as one of the very important services provided to WFCC members, which provides searching and statistics tools for culture collections or strains. Up to now, 640 international culture collections from 70 countries have been registered in WDCM.

The increasing demands on culture collections for authenticated, reliable biological material and associated information have paralleled the growth of biotechnology. However, only nearly one-sixth of collections registered in WDCM have their online catalogue, which greatly hinders the visibility and hence the accessibility of strains. Thus, WDCM started an international project called Global Catalogue of Microorganisms (GCM) to construct a data management system and a global catalogue to help organize, unveil and explore the data resources of its member collections. GCM is expected to be a robust, reliable and user-friendly system to help culture collections to manage, disseminate and share the information related to their holdings. It also provides a uniform interface for the scientific and industrial communities to access comprehensive microbial resource information. Now 56 international culture collections from 28 countries

have joined GCM, and there are around 280,000 microbial strains in GCM. By the end of 2015, we hope to have more than 100 international culture collections join GCM.

Big Data and Big Ideas

Shuichi Iwata
Professor at the Graduate School of Project Design of Tokyo University;
Editor-in-Chief of the CODATA Data Science Journal (DSJ)

The semantics of Big Data have been given by assuming models of Big Ideas. Data on Brownian motion inspired the importance of ideas rather than computational algorithms. Wiener process, Fokker-Planck and Langevin equations, Feynman-Kac formula and so on are the milestones of big ideas. Statistical fluctuations of the number of neutrons in a subcritical reactor are not noises but signals. Power spectral density measurement of neutron fluctuations can be used to evaluate the possible range of reactivity determination. In a similar way, data on heartbeats are the signals of our body and data obtained by single molecule detection techniques endorse how thermal fluctuations (noise) play a positive role in the unique operation of biological molecular machines. Seismic waves are the signals of our earth and internal frictions are measured to see dynamics of dislocations and point defects. Fluctuation 1/f is associated with art, and probably with muscle and brain. The Black-Scholes option pricing model describes Big Data on economics, which implies dynamic energy policies.

However, these traditional relations between data and ideas are evolving and becoming more diverse. The data deluge is flooding our capacity, and obliges the development of enhanced filtering filtering functions for data. Thanks to open access and open data, data are increasingly available anytime, any place and for everyone. Yet, time is limited and data literacy is limited. Is there any break-even point between increasing data and data processing efficiency? Is compression of Big Data multiplied by copy/edit/paste feasible by abstraction/selection/structuring? Does the digital divide of data usage and benefit become larger and larger due to missing ideas on being, doing and functioning?

So as to develop humankind by taking advantage of the Big Data era, this talk will make a proposal to re-engineer established disciplines through common actions from observation to design.

Session B - Big Data Today: Lessons and Challenges from Science Initiatives and the Commercial World

Genomics: Big Data, Big Science, Big Sharing

YANG Huanming
President of BGI-China

The discovery of the DNA double helix in 1953 revealed the basis of genetic information and inspired the international Human Genome Project in the 1990s. This in turn, has had three major impacts on life sciences: cultivating a *Culture of Global Collaboration*, providing an important *Technology of Sequencing*, and making a big *Science of Genomics*.

Genomics has two pillars, "*Life is of Sequence*" and "*Life is Digital*". Sequencing technology makes life digital and lays the foundation of the Big Data era of life sciences.

BGI, as one of the most influential genomics organizations in the world, has not only contributed approximately one third of extant genome sequence data, but also has been persistent in the free-sharing of genomics data, in close collaboration with its partners globally.

Datafication of the World: People, Community, and Physical Space

Eric Chang
Senior Director of Technology Strategy at Microsoft Research Asia (MSRA)

While Big Data has become a popular phrase in recent years, in this talk, I will discuss how using data better to model, interpret, and represent the world has been a continual process throughout the ages. Data allows us to build models to understand how the world functions, so that we can make better predictions and decisions and view the world from new perspectives. I will use research projects from Microsoft Research to illustrate how data provide novel and richer ways to engage with people, community, and the physical space.

Chemical, Material, and Physical Property Big Data: Balancing Scientific and Commercial Interests

John Rumble
President of R&R Data Services in Gaithersburg MD

The Twentieth Century saw the creation of material goods that revolutionized society including transportation, communication, agriculture and food, housing and furniture, clothing, and commercial and consumer products. This success was due in large part to advances in the chemical, physical, and materials

Workshop on Big Data for International Scientific Programmes

sciences. The companies that have created these revolutionary goods have succeeded, and this success in large part resulted from the widespread availability of reliable data generated by fundamental research complemented by critical proprietary data coming from commercial research and development. The importance of these data have led to a self-supporting data industry, especially for chemistry and materials science. The emerging Big Data tools and knowledge discovery methodologies are being rapidly adopted by commercial enterprises to exploit their existing data collections, but done on a proprietary basis. The power of Big Data, however, is most applicable when used with comprehensive data sets, and in these fields, many of the data sets are for-fee or proprietary. Getting access to those data sets can be expensive, difficult to do because of licensing, or impossible because of proprietary reasons. In this presentation, I will describe today's landscape for chemical, material, and physical property data as well as the features that makes creating "virtual" Big Data collections difficult. I will also discuss the consequences of these difficulties as well as possible approaches to overcome the barriers.

The Observations and Data Revolution: Opportunities for Future Earth

Frans Berkhout
Interim Director of Future Earth

Observations of many different kinds are central to understanding socio-ecological systems. Much of the effort at the international level has been on Earth Observations from space and the exchange of data collected at national level. Increasingly more powerful and accessible Earth observation platforms are becoming available, with the private sector playing a leading role in innovations. At the same time, more crowd-sourced observations are becoming available and more data is accessible on the web. Linking and combining these different forms and flows of data, about natural and social systems, is transforming the way we will do global change science, as well as offering new ways of tackling sustainability challenges. This presentation will discuss how these opportunities are being addressed by the Future Earth programme.

Policy Issues in Big Data

Paul Uhlir
Director of the Board on Research Data and Information (BRDI) at the U.S. National Academies

"Big Data" are not just quantitatively different from smaller data sets; they have characteristics that are also qualitatively different and that raise new science policy issues that need to be identified and resolved. These evolving changes affect both the public and the private sectors, and all research disciplines, albeit in different ways. The new problems may be resolved at the "soft law" policy level, or may require more formal changes in national legislation and regulation, or even possibly be considered as elements of international treaties. Any such approaches, however, must be preceded by the identification and analysis of the issues that need to be addressed and a determination of options for resolving them.

This presentation is limited to the first step—the identification of those issues that will likely require resolution through new or augmented approaches. Although many of the problems that can be identified are similar across all disciplines, there are some that are quite specific to a research area, as in the case of research data issues generally and those differences will be highlighted.

Session C - Invited Talks and Discussion - Big Data Opportunities: How to Make Progress

NIST Big Data Public Working Group & Standardization Activities

Wo Chang
NIST Data Coordinator; Digital Data Advisor for the NIST Information Technology Laboratory (ITL)

Big Data is the term used to describe the deluge of data in our networked, digitized, sensor-laden, information-driven world. There is a broad agreement among commercial, academic, and government leader/s about the remarkable potential of “Big Data” to spark innovation, fuel commerce, and drive progress. The availability of vast data resources carries the potential to answer questions previously out of reach. However there is also broad agreement on the ability of Big Data to overwhelm traditional approaches. The rate at which data volumes, speeds, and complexity are growing is outpacing scientific and technological advances in data analytics, management, transport, and more.

Despite widespread agreement on the opportunities and current limitations of Big Data, a lack of consensus on some important, fundamental questions is confusing potential users and holding back progress. How is Big Data different from the traditional data environments and related applications that we have encountered thus far? What are the essential characteristics of Big Data environments? How do these environments integrate with currently deployed architectures? What are the central scientific, technological, and standardization challenges that need to be addressed to accelerate the deployment of robust Big Data solutions?

The focus of the NIST Big Data Public Working Group (NBD-PWG) is to form a community of interest from industry, academia, and government, with the goal of developing consensus definitions, taxonomies, reference architectures, and technology roadmap which would create a vendor-neutral, technology and infrastructure agnostic framework to enable Big Data stakeholders to pick-and-choose best analytics tools to meet their requirements on the most suitable computing platform and cluster while allowing value-added from Big Data service providers.

This presentation will address the past, current, and future activities of the NBD-PWG, the collaboration work between NBD-PWG and the Research Data Alliance, and the tasks, deliverables, and the timelines for the ISO/IEC JTC 1 Study Group on Big Data.

Using Big Data Analytics for Knowledge Extraction and Collaboration

Sara Graves
Secretary General of CODATA

The collection and computational capabilities available globally have provided even larger volumes of data for the scientific research and end-user communities. In order for vast amounts of data to be usable to an increasing number of interested international collaborators, we must address the need for advanced analytics to allow for the appropriate exploitation of Big Data. This presentation will describe some of the challenges and approaches in Big Data analytics for extraction of meaningful information for international scientific collaboration. The importance of creating and maintaining sustainable software and data infrastructures, along with the role of ontological approaches with the diversity of data will also be considered.

Big Data Infrastructure and Analysis: How to Make Progress?

Pit Pichappan
Senior Scientist at the Digital Information Research Foundation (India & U.K.)

Change is the data dynamic! The capacity and know-how to compute and simulate, to extract meaning out of vast data quantities and to access scientific resources are central to the new way of co-creating knowledge. The large-scale data networks where Big Data is available need both new data analysis algorithms and a new class of systems to manage voluminous data growth. Much sophisticated data analytics can produce high processing techniques for integrating structured and unstructured data analytics. Data management issues are both managerial and technical. Managerial issues include but are not limited to generation, storage, access, metadata, rights, semantics etc; technical issues include simulation, visualization, processing, tools, software, and so on. Data infrastructure provision across countries and laboratories is skewed! Big Data technologies are still immature and evolving. Data is often managed inconsistently, stored in multiple disparate locations, and ultimately not accessible to those who need it most in time and in a form they can readily assimilate. Co-location in distributed repositories offers some equations. It is one of the most cost-effective ways of deploying Big Data clusters. Metadata at source could offer a platform to do the description of the data at source level wherein complexities can be addressed effectively. Data management needs mechanisms to show metadata relationships between data as opposed to the contents of the data itself. Clouds are also proposed as object storage repositories as we encounter multiple clouds and each cloud can be suited to one specific object storage paradigm and sharing the data with other cloud is prevented. Shared data affect sensitivity which can be solved by contemporary encryption. One of the most effective ways of integrating infrastructure lies in creating mechanisms for managing and sharing distributed repositories. We have more avenues open and have room for technology improvement to determine how we're going to share the data and how we're going to store it.

Data-Intensive Scientific Discovery in Digital Earth

WANG Lizhe
Professor at RADI, CAS

In this talk Prof. Wang will give a short introduction on his work on "Data-intensive Scientific Discovery in Digital Earth" at CAS. Prof. Wang's group mainly carries two research directions currently: 1) Signal processing and machine learning technologies are used, such as sparse representation, dictionary learning, and deep learning, to make a massive data set into a sparse data set and finally to find relativity (knowledge) in the massive data set; 2) high performance computing and cloud computing technologies are used, to develop highly efficient massive data storage, processing, scheduling and service framework.

Session D - Opportunities for Collaboration on Big Data for International Science

From TerraPop to Teraflops: The Challenge of Integrating Big Social and Natural Science Data

Robert Chen
Director of the Center for International Earth Science Information Network (CIESIN) at the Earth Institute of Columbia University

Data volumes and processing capabilities are growing rapidly in both the natural and social sciences, but not necessarily in ways that foster better integration of natural and social science data and knowledge. Concerted efforts are needed to identify points of overlap and develop meaningful connections between diverse data types. The Terra Populus project (TerraPop for short) is a focused effort to integrate social science data on billions of people around the world with a wide range of other environmental and socioeconomic datasets to support interdisciplinary research on human-environment interactions. Led by the Minnesota Population Center and funded by the U.S. National Science Foundation, TerraPop is working to address a range of technical, scientific, and ethical issues including harmonization of variables across national censuses and changing boundaries, enabling of queries across very different data types, and limitation of disclosure risk.

Data for Research That Connects Population, Agriculture and Environment: Opportunities and Challenges

Myron Gutmann
Research Professor of the Institute for Social Research at the University of Michigan

Much of the discussion of access to and use of new (and big) data for research revolves around assertions of the availability and importance of these new sources of information without focusing on the challenges they pose for end-use researchers and data scientists. In this talk I describe a long-term research agenda in both substantive analysis and data science for an emerging class of new and large resources that I call “integrated” data, which emerges from my experience as a substantive researcher, as a manager of a large data repository, and as someone responsible for social science infrastructure investments at the National Science Foundation. The starting point is my own research that integrates social and environmental science about the United States. I expand on that use case with other examples of new, large, and integrated data about population, environment, and social change. I conclude my presentation by discussing the challenges posed by new, large, and integrated data for access by researchers, as well as opportunities for future research and development in data science.

Big Data: Issues in Organizing and Retrieval

A.R.D. Prasad
Head of the Documentation Research and Training Centre (DRTC) at the Indian Statistical Institute (ISI)

One of the purported objectives of Semantic Web is to get answers from the web rather than web pages. This would distinguish the present day web from that of the future. The future web would also include a large number of data sets from various disciplines and domains. A shift in paradigm from web to data web warrants a new technology framework. The present study is centred around the agricultural data sets available on the site *http://data.gov.in*. It is a Government of India initiative to promote open data. The data sets are hosted in different file formats (csv, xml, excel, etc). However, there is no mechanism on the site to make the data discoverable effectively. In other words there exists no metadata description of the hosted data sets. Here we make an attempt not only to provide Dublin Core metadata description to make the data discoverable, but also enrich/extend Dublin core by adding a few more elements to make the data amenable to SPARQL queries. In a way we extend the goal of metadata from mere discoverability to that of making the data amenable to processing. The metadata should help in making the data more granular or even converting data on the fly to RDF. The goal is not just retrieving the tables from the site, but to answer specific and relevant queries on the tables. Firstly, the system should retrieve relevant tables amongst the available tables by verifying whether they can provide answers to the users' queries. This task requires identification of relevant tables and establishment of cross-links amongst related tables which can be achieved by adding ontology to the metadata element set. Secondly, the system should perform required tasks to answer the queries on the data. The purpose of this talk is to elaborate on the experience of working with data sets of apex research labs in India in order to mine the information and then discover knowledge, thereby achieving the goal of semantic technology.

Can We Recommend the Best Data for Researching a Certain Disaster Event? - Multidisciplinary Data Correlation Analysis as Example

LI Guoqing
Professor at RADI, CAS

There are various types of disciplinary data to which scientists are referring when they talk about the multidisciplinary data needed to support disaster research, for example geological data for earthquake study and hydrological data for flood study. However, how do we know that the right choice of data is being made? What is the scientific method to find out the right data type for a given disaster event? Big Data analysis on correlation relationship between a disaster event and the supporting multidisciplinary data has shown us a positive means of clarifying this issue. A study with key words of data citation from the published literature has been use to analyse the relationship between disaster events and the supporting data type. For disaster events with given time and position, the quantitative contribution of different disciplinary data types is sensitive. Some of such dependency changing trend of support data type can be explained by the evolution of disaster mitigation technology and policy. It may bring a mechanism to recommend scientists the right collection of multidisciplinary data to analyse for given disaster events.

Session E - Opportunities for Collaboration on Big Data for International Science

GigaDB: Open Data for the Big Data Era

Chris Hunter
Lead Bio-Curator for GigaDB at GigaScience

Traditional methods of disseminating research such as publication are bottlenecks in a system that is struggling to cope with increasing data volumes. There is also a noticeable trend of more and more publications becoming unreproducible, due to the lack of data availability, that is leading to an ever-increasing number of retractions. New platforms for disseminating data, results and methods are required to maximize knowledge discovery from these precious data resources, and these need to be made freely and conveniently available to the scientific and wider community interested in carrying out data-driven discovery.

GigaScience is an open-access, open-data journal attempting to revolutionize large-scale biological data dissemination, organization and re-use through publication. Utilizing the extensive informatics infrastructure of the BGI, the world's largest genomics organization, GigaScience links standard manuscript publication with an integrated database (*www.GigaDB.org*) that hosts all associated data and provides data analysis tools and computing resources. In addition, open-source platforms such as the popular Galaxy workflow management system are used by GigaScience to make publishing more transparent and open by making all of the supporting workflows and methods available, thereby promoting reproducibility – for which the authors are credited.

With data citation producing evidence of its use in the wider research community, GigaScience hopes to revolutionize the publication model with the aim of executable publications, where data analyses can be reproduced and built upon by users without a coding background or heavy computational infrastructure and in a more democratized manner.

Efficiency of Knowledge Extraction on the Big Data Framework

Jan-Ming Ho
Research Fellow at the Institute of Information Science, Academia Sinica

Information technology has been widely adopted in all sectors of industry and business, and our daily lives. Streams of huge amount of data have thus been accumulated more rapidly than ever. In response to the Big Data challenge, a spectrum of open source software packages including the MapReduce framework has become available to help the scientific community derive knowledge from data. In this talk, we will present our recent efforts in studying the computing structures of NGS genome sequence assembly based on the MapReduce framework. Emphasis will be made on sequence error correction as an example to illustrate the importance of designing efficient and effective MapReduce algorithms.

Force Multiplier for Science through Crowd Source Validation and Visualization

Tim Foresman
SIBA Chair for Spatial Information at the Queensland University of Technology

Increasing sources of data are rapidly adding to what is generally referred to as Big Data, or BD henceforth. Estimates by IBM researchers indicate that 90% of the world's data in existence was created within the past two years. Furthermore, it is generally acknowledged that approximately 85% of these data are inherently spatial, which means that it can all be interrelated within a Digital Earth framework. To effectively and authoritatively apply these data requires use, however, of protocols and standards that have been carefully identified and codified over the past few years. Currently, the advent of crowd sourcing has been viewed as a force multiplier for many areas of science and governance adding challenges for information analysts operating in the BD arena. This presentation articulates collaborative efforts within the ISDE and Australian communities to address the potential of crowd sourcing and to enhance citizen engagement for improved monitoring and sustainable management of our Earth's systems.

BD dialogues have popularly accepted the four Vs to describe characteristics of explosive data creation: volume, velocity, variety, and veracity. Crowd source (also known as volunteer geographic information (VGI)) is similar to citizen science with important distinctions. Citizen science entails the directed collection, following general guidelines from subject matter experts (SMEs) to augment existing and ongoing data collection efforts. Cornell's bird census and OpenStreetMap are examples of well-tested schemas. Crowd sourcing, which owes its etiology to Surowiecki and Galton, applies less rigorous upwelling of citizen-originated information into both organized (e.g., Ushahidi, Tomnod) and mass aggregations of data

residing in server farms or clouds (e.g., Facebook, Twitter). Effective application of this information resides in (1) determining the veracity of the source information and (2) visualizing the data trends for analysis, which drives our research focus.

Global Change Research Data Publishing and Repository in China

LIU Chuang
Professor at the Institute of Geographic Sciences and Natural Resources Research (IGSNRR), CAS

Many thousands of global change research data sets have been archived in China since the Scientific Data Sharing Program launched in China in 2003. However, none of them has been identified as computer searchable by citation function. Moreover, none of the academic institutes and universities in China credits database creators for their research achievements in the same way as they credit published research papers or patents; and this despite many calls for database creation to be credited. Data publishing is the most important way to handle these problems. The initiative of the Global Change Research Data Publishing and Repository is a joint effort from the Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences (IGSNRR/CAS), the Geographical Society of China (GSC), the National Remote Sensing Center, Ministry of Science and Technology of China (NRSC/MOST), and coordinated with the CODATA Task Group of Preservation of and Open Access to S&T Data in Developing Countries (CODATA-TA-PASTD) as well as the Digital LIN Chao Geomuseum and the China GEO Secretariat. In this initiative many issues are identified and examines, including: the practices DOI registration for datasets, the responsibilities of dataset authors and editorial board members, the flowchart of the data publishing procedure, the steps of dataset peer review, and quality standards for data papers both in Chinese and English, citation between datasets and journal publications, the responsibilities of the digital library for the long term preservation of datasets, as well as online data services. More than ten organizations in China are preparing to re-organize the archived data for publishing, and this will establish cooperation between them on global change research data publishing. Global change research data publishing will great help China as a new milestone in enhancing global change data sharing, for which many Chinese scientists have been working hard since 1994.

Workshop on Big Data for International Scientific Programmes

IV. Short Biography of Speakers

Frans Berkhout

Frans Berkhout is Professor of Environment, Society and Climate in the Department of Geography at King's College London, and Interim Director of the Future Earth programme, based at the International Council for Science (ICSU) in Paris. Between 2004 and 2012, Prof. Berkhout was director of the Institute for Environmental Studies (IVM) at the VU University Amsterdam in The Netherlands, and from 2010 to 2013 director of the Amsterdam Global Change Institute. Professor Berkhout is a lead author in the IPCC's Fifth Assessment Report (2014) and a member of the Research Evaluation Framework (REF) of the Higher Education Funding Council for England (HEFCE). He sits on the editorial boards of *Research Policy*, *Global Environmental Change*, *the Journal of Industrial Ecology*, *Current Opinion on Environmental Sustainability* and *The Anthropocene Review*. His early research was concerned with the economic, political and security aspects of the nuclear fuel cycle and radioactive waste management. His more recent work has been concerned with science, technology, policy and sustainability, with a focus on climate change.

Eric Chang

Dr. Eric Chang joined Microsoft Research Asia (MSRA) in July 1999 to work in the area of speech technologies. Eric is currently the Senior Director of Technology Strategy at MSR Asia, where his responsibilities include communications, IP portfolio management, and driving new research themes such as eHealth and urban informatics. Prior to joining Microsoft, Eric had worked in Nuance Communications, MIT Lincoln Laboratory, Toshiba, and General Electric Corporate Research and Development. Eric graduated from MIT with Ph.D., Master and Bachelor degrees, all in the field of electrical engineering and computer science. Eric's work has been reported by *Wall Street Journal, Technology Review*, and other publications.

Wo Chang

Mr. Wo Chang is NIST Data Coordinator and Digital Data Advisor for the NIST Information Technology Laboratory (ITL). His responsibilities include working with the NIST Scientific Data Committee to respond to data requirements; promoting a vital and growing Big Data community at NIST and with external stakeholders in the commercial, academic, and government sectors; and interacting with the Research Data Alliance for standards related data development. Mr. Chang is currently the Convener of the ISO/IEC JTC 1 Study Group on Big Data, he co-chairs the NIST Big Data Public Working Group, and chairs the ISO/IEC JTC/1 SC 29 WG11 (MPEG) Multimedia Preservation AHG.

Prior to joining ITL Office, Mr. Chang was manager of the Digital Media Group in ITL and his duties included overseeing several key projects including digital data archiving and preservation, management of electronic health records, motion image quality, cloud computing, and multimedia standards. In the past, Chang was the Deputy Chair for the US INCITS L3.1, chaired several other key projects for MPEG, participated with the HL7 and ISO/IEC TC215 for health informatics, IETF for the protocols development, and was one of the original members of the W3C's SMIL and developed one of the SMIL reference software.

Deliang Chen

Deliang Chen is August Chair at the University of Gothenburg, Sweden. As an internationally renowned climate scientist, he worked as Lead Author for WG1 of IPCC AR5, and has made an important contribution in the area of global change focusing on regional climate changes in Sweden and China. During 2009-2012 he served as Executive Director of the International Council for Science (ICSU). He has served on numerous international and national science committees and boards for international research programmes, national funding agencies, and well-known research centers. Examples include Chair of the Scientific Advisory Committee for Environment Climate Data Sweden, Science Director of the Beijing Climate Center, Member of the French ANR Scientific Steering Committee on Earth System Science, Member of the Steering Committee for World Science Forum Series, Member of the Advisory Group for the OECD Programme on Innovation, Higher Education and Research for Development, Member of the Oversea Expert Advisory Committee of the State Council of China, Member of the Project Evaluation Committee of the Japanese Research Institute for Humanity and Nature (RIHN), and Board Member of the Stockholm Resilience Centre. He also serves in various selection committees for well-known international science awards such as the Volvo Environment Prize.

Workshop on Big Data for International Scientific Programmes

Robert Chen

Robert Chen is Director of CIESIN, the Center for International Earth Science Information Network, a research unit of the Earth Institute at Columbia University in New York. He is a member of the U.S. National Research Council (NRC) Board on International Scientific Organizations, the Board of Directors of the National Ecological Observatory Network (NEON), and the Governing Council of the Interuniversity Consortium for Political and Social Research (ICPSR). He co-manages the Intergovernmental Panel on Climate Change (IPCC) Data Distribution Center and is a member of the Scientific Leadership Team of Terra Populus, a DataNet project led by the University of Minnesota. From 2004-12, he served as Secretary General of the Committee on Data for Science and Technology (CODATA) of the International Council for Science (ICSU). At Columbia, Dr. Chen is an ex officio member of the Earth Institute Faculty and a member of the Faculty Steering Committee for the Columbia Global Centers | East Asia. His research areas include data access and stewardship, disaster risk assessment, and climate change adaptation and indicators. He received his Ph.D. in geography from the University of North Carolina at Chapel Hill and holds B.S. and M.S. degrees from the Massachusetts Institute of Technology.

Tim Foresman

Tim Foresman is a senior science, engineering, and education leader. He currently serves as the SIBA Chair for Spatial Information at Queensland University of Technology, were he is responsible for research and teaching of technology and applications associated with the emergent and rapid growth of the spatial information phenomena. He has been active in helping organizations reach their full potential and address the challenging issues of better stewardship for managing human and ecological resources while focusing on the goals for sustainable and higher living standards. Previously, he served as the United Nations' chief environmental scientist with the United Nations Environment Programme, Nairobi, Kenya. He has been a technology leader in the use of scientific visualization and spatial information systems (remote sensing and GIS) for community decision support (*www.earthparty.org*). He served NASA Headquarters as the national manager for the Digital Earth Initiative under Vice President Al Gore, which introduced many developments, including Google Earth, and the establishment of the International Society for Digital Earth (ISDE) with a Beijing, China secretariat. He is a founding member of ISDE.

Peter Fox

Peter Fox is Tetherless World Constellation Chair, Professor of Earth and Environmental Science, Computer Science and Cognitive Science, and Director of the Information Technology and Web Science Program at Rensselaer Polytechnic Institute. Fox has a B.Sc. (hons) and Ph.D. in Applied Mathematics (physics and computer science) from Monash Univsersity. His research covers the fields of computational and computer science, ocean and environmental informatics, distributed semantic data frameworks and solar and solar-terrestrial physics. The results are applied to large-scale distributed data science investigations. Fox is President of the Federation of Earth Science Information Partners (ESIP), chair of the International Union of Geodesy and Geophysics Union Commission on Data and Information, and serves on the editorial boards of many prominent Earth and space science informatics journals. In 2012, Fox was awarded the European Geoscience Union, Ian McHarg / Earth and Space Science Informatics Medal, and ESIP's Martha Maiden Lifetime Achievement award for service to the Earth Sciences Information communities.

Sara Graves

Dr. Sara James Graves is the Director of the Information Technology and Systems Center, the University of Alabama System Board of Trustees University Professor and Professor of Computer Science at the University of Alabama in Huntsville.

Her current service includes the Secretary General of CODATA, the International Council for Science: Committee on Data for Science and Technology; the National Academy of Sciences Board on Research Data and Information; the Science Advisory Board for the Oak Ridge Climate Change Science Institute; the Board of Trustees for the Southeastern Universities Research Association (SURA); and the Executive Committee of the Earth Science Information Partners Federation.

Dr. Graves directs research and development in Big Data analytics and visualization, semantic technologies, data mining and knowledge discovery, and sustainable software infrastructures. Her degrees are in Computer Science and Mathematical Sciences, and she has served as chair or member of over 100 Ph.D. and M.S. committees.

GUO Huadong

Prof. Guo Huadong is Director-General of the Chinese Academy of Sciences (CAS) Institute of Remote Sensing and Digital Earth (RADI), an Academician of CAS, and a Fellow of the Academy of Sciences for the Developing World (TWAS). He presently serves as President of the International Council for Science (ICSU) Committee on Data for Science and Technology (CODATA), Scientific Committee Member of the Integrated Research on Disaster Risk (IRDR) programme, and Editor-in-Chief of the *International Journal of Digital Earth* (IJDE) published by Taylor & Francis. He has over 30 years of experience in remote sensing, specializing in radar for Earth observation and remote sensing applications, and has been involved in research on digital Earth since the end of the last century. He has been Principle Investigator for over twenty major national projects or programs in China, and Principle Investigator for seven international radar remote sensing projects. He also serves as Director of the International Center on Space Technologies for Natural and Cultural Heritage under the Auspices of UNESCO. Prof. Guo has published more than 400 papers and 16 books, and is the principal awardee of 13 national and CAS prizes.

Myron Gutmann

Myron P. Gutmann is Professor of History and Information and Research Professor in the Institute for Social Research at the University of Michigan. From 2009 to 2013 he served as Assistant Director of the U.S. National Science Foundation, leading NSF's Social, Behavioral, and Economic Sciences Directorate. Gutmann has broad interests in interdisciplinary research, especially health, population, economy, energy, and the environment. Previously, he was Director of the Inter-university Consortium for Political and Social Research (ICPSR), the world's largest repository of publicly available data in the social and behavioral sciences. As director of ICPSR, he was a leader in the archiving and dissemination of electronic research materials related to society, population, and health, with a special interest in the protection of respondent confidentiality. At NSF Gutmann led efforts in scientific integration, ensuring that the human aspects of scientific issues are always among the questions and investments in NSF-supported science. He also spearheaded NSF's initiative to improve access to publications and data. Gutmann has written or edited five books and more than eighty articles and chapters, and has served on numerous advisory committees and editorial boards. He is an elected fellow of the American Association for the Advancement of Science.

Jan-Ming Ho

Jan-Ming Ho received his Ph.D. degree in electrical engineering and computer science from Northwestern University in 1989. He received his M.S. at Institute of Electronics of National Chiao Tung University in 1980 and his B.S. in electrical engineering from National Cheng Kung University in 1978. Dr. Ho joined the Institute of Information Science, Academia Sinica as an Associate Research Fellow in 1989, and was promoted to Research Fellow in 1994. In 2000-2003, he served as Deputy Director of the institute. In 2004-2006, he had served as Director General of the Division of Planning and Evaluation, National Science Council. He visited IBM's T. J. Watson Research Center in summer 1987 and summer 1988, the Leonardo Fibonacci Institute for the Foundations of Computer Science, Italy, in summer 1992, and the Dagstuhl Seminar on Applied Combinatorial Methods in VLSI/CAD, Germany, in 1993.

Dr. Ho's research interests cover the integration of theory and applications, including combinatorial optimization, information retrieval and extraction, multimedia network protocols and their applications, bioinformatics, open source, web services, and digital library and archive technologies. Dr. Ho has also published results in the field of VLSI/CAD physical design.

Chris Hunter

After completing undergraduate studies in applied biology, Dr. Hunter undertook a Ph.D. with Darwin College, Cambridge University on genome evolution. Over the following post-doctoral positions at the MRC-HGMP-RC and Wellcome Trust Sanger Institutes, Dr. Hunter migrated away from "wet-lab" biology finding his strength lay in data organization and manipulation rather than data acquisition. After that he spent five years at the world-renowned European Bioinformatics Institute (EMBL-EBI) specializing in DNA databases and Metagenomics analysis, being a founding member of the European Metagenomics team at EBI (*www.ebi.ac.uk/metagenomics*). Dr. Hunter joined the GigaScience team in April 2013 as the lead Bio-Curator for GigaDB (*www.gigadb.org*), the data-warehouse behind GigaScience publications.

Shuichi Iwata

Born on January 29, 1948 in Chiba Prefecture, Japan. Currently works for "Data and Society" - Making "Data on Science and Technology" available for everyone.

He became Doctor of Engineering in 1975. He was appointed Lecturer by the Department of Nuclear Engineering, the University of Tokyo between 1978 and 1991, subsequently working as Associate Professor of Metallurgy Division, Engineering Research Institute, and Department of Nuclear Engineering, University of Tokyo/Guest Researcher, Fachinformations Zentrum, BRG (1985-1986). In 1991 he became Professor of Design Science, Life Cycle Engineering and Director of RACE, Graduate Schools of Engineering. President of CODATA, Committee on Data for Science and Technology (2002-2006); Professor at Graduate School of Frontier Sciences in the University of Tokyo (2004-2012), Guest Researcher of National Museum of Nature and Science (2006-2012), Professor at Graduate School of Project Design (2012-present). Member of Science Council of Japan, Member of Engineering Academy of Japan. Editor-in-Chief: *CODATA Data Science Journal.*

Awards and honors: Honda Memorial Young Researcher Award; Iketani Science Foundation Award; Promotion of Science and Technology Information Award, Japan Science and Technology Agency; Paper Award, The Japan Institute of Metals; GIW Best Paper Award.

LI Guoqing

Dr. Li Guoqing is a Professor of the Institute of Remote Sensing and Digital Earth (RADI) at Chinese Academy of Sciences. His research now mostly focuses on next generation spatial data infrastructure and nature disaster data management.

He is the senior member of IEEE and High Performance Computation Society in China Computer Federation (CCF), as well as the director member of Chinese Specialty Association of Mathematical and Scientific Software. He has been the coordinator of dozens of National and International research projects. Dr. Li is now serving as a Scientific Committee member of the ICSU World Data System (ICSU-WDS), and as coordinator of the ICSU-CODATA LODGD Task Group. He also has been involved in the Chinese national committees of ICSU, as well as CODATA-China and IRDR-China. He has been the member of CEOS/WGISS since 2003, and served as the Grid Task Team leader and vice-chair of WGISS Application Subgroup.

LIU Chuang

Dr. Liu Chuang, Professor of Institute of Geographical Sciences and Natural Resource Research, Chinese Academy of Sciences (IGSNRR/CAS), is Director of Executive Committee of the Digital LIN Chao Geomuseum launched by the International Geographical Union (IGU), CODATA and the Geographical Society of China. She is Vice Chair of the Expert Group of China National Committee for GEO and Secretary of the CODATA Task Group on Preservation of and Open Access to Scientific and Technical Data in/for/with Developing Countries (CODATA-PASTD). She got her Ph.D. degree in geography from Peking University, China, in 1989. She was Visiting Professor at the University of British Columbia, Vancouver, Canada from 1992-1993 and Information Scientist of the Center of International Earth Science Information Network (CIESIN), USA (1994-1998). She was Director of the Global Change Information and Research Center, IGSNRR/CAS (2000-2010), and Director of World Resources Research of Beijing Normal University (2006-2010). Prof. Liu's achievements were recognized by several awards, including Best Post Award by ISPRS, 1988 in Japan; Outstanding Award of Land Management Bureau of China in 1994; National Achievement Award of China in 1996; CIESIN Achievement Award in 1996, USA; and the CODATA Prize in 2008.

MA Juncai

Dr. Ma Juncai is the director of information center of the Institute of Microorganisms at Chinese Academy of Sciences; he is also the director of WFCC-MIRCEN World Data Center of Microorganisms (WDCM), and co-chair of the Data Mirror Working Group of International DNA barcode of Life Project (iBOL), and co-chair of the CODATA Task Group on Advancing Informatics for Microbiology (TG-AIM). Dr. Ma is also in charge of the Biotechnology Information Network and the Industry Biotechnology Information Network in China.

Dr. Ma received his Ph.D. degree from the Department of Bio-resources of Mie University in Japan in 2006. As a senior visiting scholar, he worked in IBM Japan Company in 1985, American Type Culture Collection (ATCC) in 1992, and Japan Collection of Microorganisms (JCM) of RIKEN in 1995, Dr. Ma gained Academic Award and Outstanding Young Scientists Fellow from that Chinese Academy of Sciences in 1989, and National Academic Award in 1992.

Pit Pichappan

Professor Pit Pichappan is currently the Senior Scientist at the Digital Information Research Foundation in India and UK. He obtained his Ph.D. in 1994 in Scientometrics. He has served as Professor of Information Systems in many universities. He specialized in research evaluation, data mining, data processing and text mining. He has more than 100 research publications in many reviewed journals. He is associated with the editorial board of 11 journals. He is a member of the IEEE Committee on Internet Computing, Internet Society and many other bodies. He was nominated as the most eminent national expert to the World Summit Award, Geneva. He was the keynote speaker in many international conferences. He handled many funded research projects across countries.

A.R.D. Prasad

Prof. A.R.D. Prasad is the Head of Documentation Research and Training Centre (DRTC), Indian Statistical Institute (ISI), Bangalore, India. He serves on several international committees including the G8+06 committee for Data Science and Data Infrastructures, Evaluation Committee for grants of European Commission research projects. He is a consultant to UNFAO, Rome. He is a principal co-investigator of European Commission projects, 'Living Knowledge' and 'AgINFRA'. He is the scientific lead of the India-Trento Programme for Advanced Research (ITPAR).

Prof. A.R.D. Prasad is presently a member of the high level committee of the Indian National Mission for Libraries and he also served as a member of National Knowledge Commission of India. He was member of the DSpace Governance Advisory Board, Massachusetts Institute of Technology. He served as a member of several committees including University Grants Commission (UGC) Curriculum Development Committee, UGC National committee on ETDs (Electronic Theses and Dissertations) and the National Library of India, Retro-conversion committee. Author of the 'Open Mantra for Open Access', he is a staunch supporter of Open Science and Open Data.

Jakob Rhyner

Jakob Rhyner holds a Diploma and Ph.D. in Theoretical Physics from the Swiss Federal Institute of Technology (ETH) in Zurich.
Since 2010, United Nations University, Vice Rector in Europe and Director of the Environment and Human Security (UNU-EHS).
Professor at the Agricultural Faculty of the University of Bonn. Research focus and responsibilities: Environmental risks research and capacity building. 2012-13 Co-Chair of the Future Earth Implementation Board.

2001-2011, Swiss Federal Institute for Snow and Avalanche Research (SLF) in Davos Switzerland. Until 2006 Division Leader Avalanche Warning and Risk Management, from 2006 Institute Leader.
Responsibilities: Operational warning systems, risk management research and education.

1988-2001 ABB Corporate Research, industrial physics, project and team leader. Research focus: Physical and numerical modeling of electric power systems and materials.

John Rumble

Dr. John Rumble has long been a leader in developing and providing access to databases in a numerous S&T disciplines. From 1980 through 2004, he worked in the NIST Standard Reference Data Program. From 2004 to 2011, he was Executive Vice President of Information International Associates in Oak Ridge, Tennessee. Today he is President of R&R Data Services in Gaithersburg MD.

Dr. Rumble was among the first to build online, PC, and Internet/web-based factual databases for scientific and technical (S&T) data. During that time, he also worked to develop industry-related data programs in materials science and engineering, as well as standards for materials data exchange.

Dr. Rumble has considerable chemistry, physics, materials science and information science experience. He has published widely in all these areas. He is presently Chair of an international, multi-disciplinary Working Group on Nanomaterials. Rumble received a Ph.D. in Chemical Physics from Indiana University.

In 1998-2002, Rumble served as President of CODATA. Dr. Rumble is Fellow of IUPAC, AAAS, ASTM, and ASM International, as well as a Foreign Member of the Russian Academy of Metrology. He was awarded the CODATA 2006 Prize. He is now serving a second term as Editor-in-Chief of the *CODATA Data Science Journal*.

Barbara Ryan

Barbara J. Ryan is Secretariat Director of the intergovernmental Group on Earth Observations (GEO) located in Geneva, Switzerland. In this capacity, she leads the Secretariat in coordinating the activities of nearly 90 Member States and 50 Participating Organizations which are striving to integrate Earth observations so that informed decisions can be made across nine Societal Benefit Areas including agriculture, biodiversity, climate, ecosystems, energy, disasters, health, water and weather. Before assuming this position in July 2012, she was the Director of the World Meteorological Organization (WMO) Space Programme. She had responsibility for the space-based component of the WMO Global Observing System (GOS), coordinated space-based assets to meet the needs of WMO Members in the topical areas of weather, water, climate and related natural disasters, and also served as the technical focal point for WMO's activities with GEO. Before joining WMO in October 2008, she was the Associate Director for Geography at the U.S. Geological Survey (USGS) in Reston, Virginia where she had responsibility for the Landsat, remote sensing, geography and civilian mapping programs of the agency. It was under her leadership that implementation of the Landsat data policy was reformed to release all data over the internet at no additional cost to the user – an action that has resulted in the release of more than 12 million Landsat scenes to date. As the 2007 Chair of the international Committee on Earth Observation Satellites (CEOS) she led the space-agency response to the Global Climate Observing System (GCOS) satellite requirements for sustained measurement of the GCOS Essential Climate Variables (ECVs). She holds a Bachelor´s degree in Geology from the State University of New York at Cortland, a Master´s degree in Geography from the University of Denver, and a Master´s degree in Civil Engineering from Stanford University.

Mark Stafford Smith

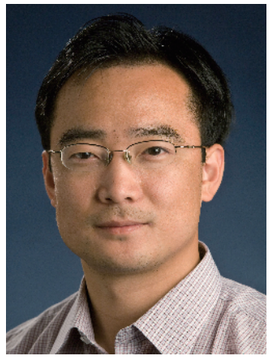
Dr. Mark Stafford Smith is the Science Director of CSIRO's Climate Adaptation Flagship in Canberra, Australia, where he oversees a highly interdisciplinary program of research on many aspects of adapting to climate change, as well as regularly interacting with national and international policy issues. He has over 35 years experience in drylands systems ecology, management and policy, including senior roles such as CEO of the Desert Knowledge Cooperative Research Centre in Alice Springs. His significant international roles include being past vice-chair of the International Geosphere-Biosphere Programme's Scientific Committee. In 2012 he was co-chair of the Planet Under Pressure: New Knowledge Towards Solutions conference on global environmental change in the lead up to Rio+20. In 2013 he was appointed Chair of the inaugural Future Earth Science Committee, which aims to help coordinate global change research worldwide.

Paul Uhlir

PAUL F. UHLIR is Director of the Board on Research Data and Information (BRDI) at the U.S. National Academies in Washington, DC. Paul's area of emphasis is on issues at the interface of science, technology, and law, with primary focus on digital data and information policy and management. He also directs the U.S. Committee on Data for Science and Technology. Between 1985 and 2008, Paul held a series of senior positions at the National Academies. Before that, he worked at the Office of the General Counsel and was a foreign affairs officer at the National Oceanic and Atmospheric Administration, where he focused on remote sensing law and policy and on intergovernmental agreements for cooperation in meteorological satellite programs. Paul is the author or editor of 24 books, and over 70 technical articles. He has been involved in numerous consulting and pro bono activities, and speaks worldwide on a broad range of information policy and management issues. In 1997 he received the National Research Council's Special Achievement award and in 2010 the CODATA Prize for his work on international data policy. Paul has a J.D. and an M.A. degree in international relations from the University of San Diego, and a B.A in history from the University of Oregon.

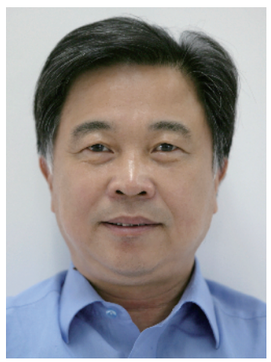
Workshop on Big Data for International Scientific Programmes

WANG Lizhe



Dr. Wang Lizhe is a Professor at the Institute of Remote Sensing & Digital Earth (RADI), Chinese Academy of Sciences (CAS) and a "ChuTian" Chair Professor at School of Computer Science, China Univ. of Geosciences (CUG). Before his tenure at CAS, he was a research scientist and principal research engineer at Indiana University, Bloomington. Prof. Wang received his B.E. and M.E. from Tsinghua Univ. and Doctor of Eng. from Univ. Karlsruhe, Germany. Prof. Wang is a Fellow of IET, Fellow of British Computer Society, and Senior Member of IEEE. He is an associate Editor of *IEEE Transaction on Computers, IEEE Transactions on Cloud Computing*, and *Journal of Supercomputing, Journal of Cluster Computing*. Prof. Wang leads a group at CAS on HPC & data-intensive computing. More information can be found at his homepage: *http://www.escience.cn/people/lzwangEN/.*

YANG Huanming

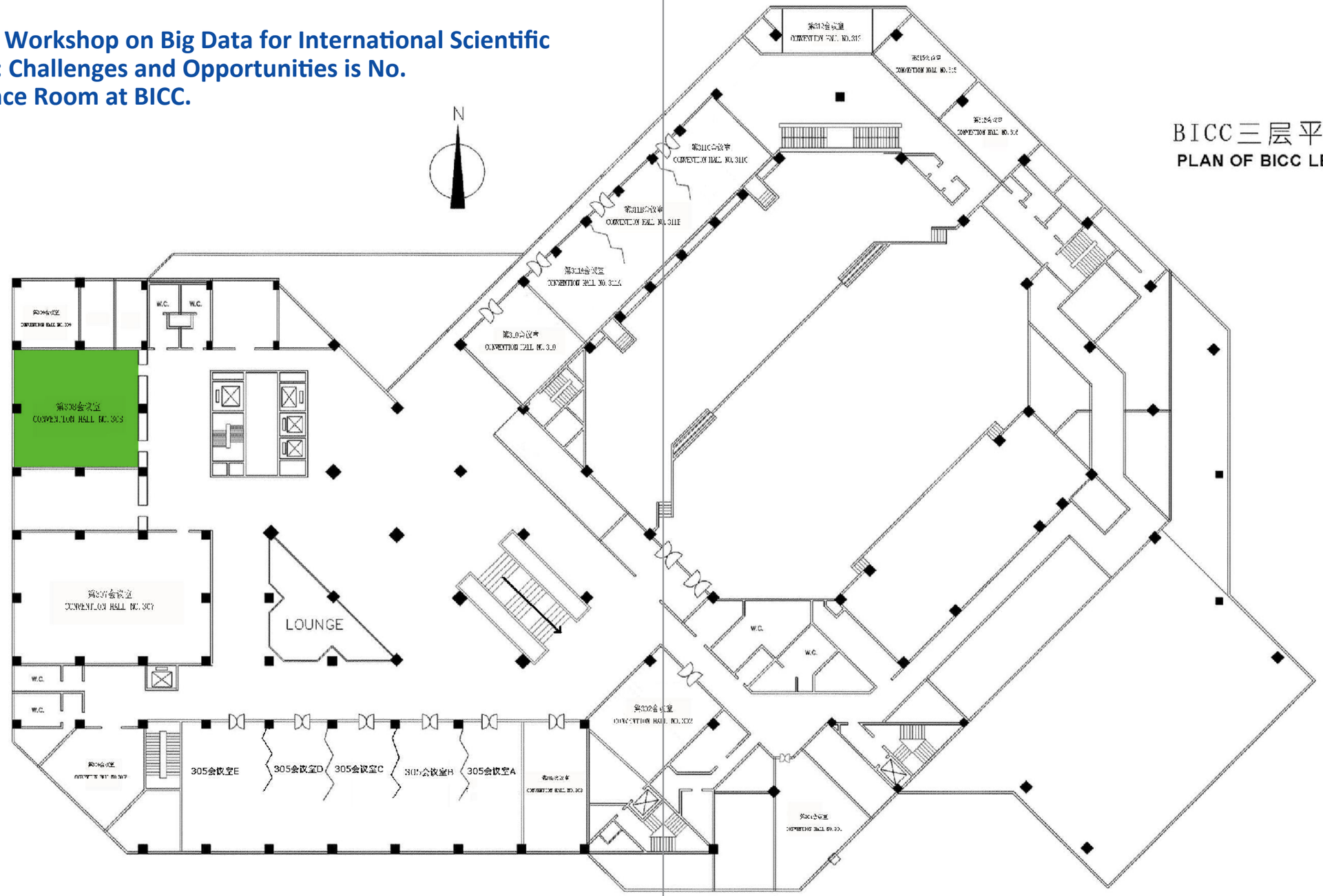


Dr. Yang is the co-founder and President of BGI-China, one of the major genomics centers in the world. He and his partners have made a significant contribution to the International HGP, HapMap Project, and other human-omics research, as well as sequencing and analyzing genomes of many other animals, plants, and microorganisms, with many publications in *Science*, *Nature*, and other internationally prestigious journals.

Dr. Yang obtained his Ph.D. from University of Copenhagen (Denmark) and postdoctoral training in France and USA. He was elected as an associate member of European Molecular Biology Organization (EMBO) in 2006, an academician of Chinese Academy of Sciences in 2007, a fellow of TWAS in 2008, a foreign associate of National Academies of India in 2009, of Germany in 2012, and of the USA in 2014.

V. Conference Venue Layout

The venue of Workshop on Big Data for International Scientific Programmes: Challenges and Opportunities is No. 308 Conference Room at BICC.



VI. General Information

About Beijing

Beijing, located in the north, is China's political centre and a busy capital city with a population of over 15 million people. As the seat of power of Chinese emperors throughout the centuries, Beijing is steeped in history, including the 26 traditions of the Ming and Qing Dynasties (1368-1911). Reigning as both an ancient capital of Imperial China and the modern capital of a thriving nation, Beijing retains plenty of evidence of its royal past, with aristocratic parks, temples, and palaces (all open to the public). Beijing is home to an incredible cultural display of art and historical artifacts in more than 50 museums. Folk traditions flourish in theaters, delicious dining is available in exotic settings, and cultural centers with fascinating demonstrations of centuries-old art and craft making abound. Nowhere else can you get a more concentrated impression of the old and new China. Beijing is the treasure trove of Chinese culture, where many of the sights that make china a world-class destination are located.

For more on the history of Beijing:

http://china.org.cn/english/features/beijing/30785.htm

http://en.wikipedia.org/wiki/History_of_beijing

About the Workshop Venue

The Beijing International Convention Center (BICC) is located in the flourishing Yayuncun area along Beijing's North Fourth Ring Road, where the central axis of the city meets the Fourth Ring Road, and right next to national stadiums like the Bird's Nest and the Water Cube. It's a 20km trip east to the airport, a 9km journey south to Tian'anmen Square, a 10km excursion west to the Summer Palace, and an 80km sojourn north to the Badaling section of the Great Wall. And with the Olympic Village only a stone's throw away, there is no better location in the city from which to base your business trip.

Recommended Hotel

The Beijing North Star Continental Grand Hotel is a large 4-star hotel with 538 elegant, comfortable rooms. The guest rooms are equipped with the latest, advanced technology which demonstrates the international quality and hospitality. During the 2008 Beijing Summer Olympic Games and the Paralympics Games this 13 story hotel with a total area of 42,000 square meters was the hotel of choice because of its full range of facilities and magnificent flowing style. As soon as you enter the hotel you are able to feel as if you are home.

VII. Acknowledgements

CODATA, the ICSU Committee on Data for Science and Technology, through its Executive Committee, would like to thank the Institute of Remote Sensing and Digital Earth (RADI) of the Chinese Academy of Sciences (CAS) for support in convening this workshop. Warm thanks are also due to the co-sponsoring organizations and to all of the speakers and participants. It is hoped that this workshop will lead to further rich collaboration and exchanges of ideas that benefit international research in the Age of Big Data.

Taking this opportunity, CODATA would like to thank Simon Hodson, LIU Jie, WANG Changlin, LIANG Dong, CHEN Mingmei, LIU Jingna, GUO Song, GUAN Linlin, LI Jiani, JIANG Hao and LU Yiqun for their hard work and tireless contribution to the organization of this workshop.