{No.4　　Vol. 139}

Shanghai Center for Pujiang Innovation Forum　　June.16 2021

2021 Pujiang Innovation Forum Bulletin IV

Scientific Data Management, Sharing and Application

Editor's Note: In Data Creates Universe – Scientific Data Innovation Conference of 2021 Pujiang Innovation Forum – The Emerging Technology Forum, centering on the innovative concepts and practices of scientific data management and sharing and application services, well-known experts and scholars at home and abroad shared excellent achievements in scientific data opening, sharing, application, interaction and collaboration, and had in-depth discussions on the policies, principles and systems of data opening. This bulletin is a summary based on the reports from the participating guests[1], and is intended for reference.

[1] Network Information Center, Chinese Academy of Sciences; JIANG Lulu, Product Manager of Science DB, Computer Network Information Center of the Chinese Academy of Sciences; Barend Mons, GO FAIR International Support and Coordination Office, President of CODATA, and Professor of Leiden University; Gergely Sipos, Director of Solution, EGI (European Grid Infrastructure)

2021 Pujiang Innovation Forum Bulletin IV

Scientific Data Management, Sharing and Application

In recent years, the development of science and technology has showed evident trends towards big science and quantitative research. Science and technology innovation increasingly depends on large number of systems, highly-reliable scientific data and comprehensive analysis and mining of the scientific data. **The participating guests pointed out that it's of great importance and an urgent need to build a network and a mechanism for global scientific data sharing, which is the key to realizing the deeper data value and building the innovation ecology of scientific data.**

I. Scientific Data Has Become a New Engine for the Development of Science and Technology Innovation

Firstly, data is the scientific key to solving complicated problems. As scientific research steps into the data-intensive "Fourth Paradigm Era", many problems will remain unsolved without data. According to the introduction provided by **GUO Huadong, Academician of Chinese Academy of Sciences, and Director and Research Fellow of the Academic Committee, Aerospace Information Research Institute, Chinese Academy of Sciences**, 41% of the 17 United Nations Sustainable Development Goals for global revolution "have approaches to success but no data". As pointed out by **XU Ren, Deputy Director of**

Polar Research Institute of China, from the signing of the *United Nations Convention on the Law of the Sea* for the Arctic Sea Route and the project approval of the new station on Antarctic's Ross Sea to the R&D of polar icebreaker "Xuelong 2", all the aforementioned are based on a huge amount of scientific expedition data.

Secondly, data empowerment has unleashed immense potential in several scenarios. In the era of big data, data empowerment is accelerating science and technology innovation, and has brought about significant pull effect, amplification effect and multiplier effect. As pointed put by **FENG Jianfeng, Dean of School of Data Science, Fudan University and Dean of Institute of Science and Technology for Brain-inspired Intelligence, Fudan University**, big biological data is the foundation of smart healthcare, of which the ultimate goal is to accurately predict individuals' physical and mental health conditions. Currently, FENG's research team could identify and recognize depression through gait (with an accuracy of over 70%) and precisely judge whether thrombolytic therapy is applicable by the patient's brain image based on the software system it developed, which breaks through the traditional diagnosis and treatment barriers to precisely recognizing the time of stroke onset. According to **Academician GUO Huadong**, his research team has proved with data that China is the biggest contributor to zero net global land deterioration, and found that the global glacial reserves reduced by 6% from 1999 to 2018, equivalent to 12mm global sea level rise.

II. Data Ecology Creates and Enables Scientific and Technological Achievements to Yield New Impetuses

Firstly, the data ecology comprehensible to both humans and machines is the key to future development. As stressed by **LI Jianhui, Engineer of Computer Network Information Center, Chinese Academy of Sciences**, with the technological development in the future, machines' capability for automatic data acquisition and interpretation will become the key to promoting the efficiency of data application. We shall cross disciplinary boundaries to establish the data ecology which could be comprehended and operated by both humans and machines. In the opinion of **Professor Barend Mons, GO FAIR International Support and Coordination Office, President of CODATA**, besides the four original principles (data are available for discovery, access, interaction and utilization), free application of data by AI shall also be included in Fair (FAIR data principles) at the current stage.

Secondly, fully-automated data processing is where we head for. According to **SUN Yangang, Deputy Director of Institute of Neuroscience, Chinese Academy of Sciences Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences**, currently, neuron reconstruction is mainly conducted in the manual or semi-automatic operation mode which is time-consuming and labor-consuming. The research into the drawing of the whole brain mesoscopic neural connection atlas of mice and rhesus monkeys, which involves a huge amount of data and extremely complicated structures, is in urgent need of three-dimensional data identification through more in-depth applications of emerging technologies including AI Deep Learning, to realize the fully-automated reconstruction of neurons. According to **Dean FENG Jianfeng**, it's hopeful to achieve high-value breakthroughs as cerebrovascular image analysis based on big brain science data has shown great application potential in smart diagnosis and treatment of traditional diseases such as depression and autism.

Thirdly, the intelligent data integration service system is the cornerstone of the construction of the data sharing system. Since the release of the *Measures for the Management of Scientific Data*, 20 national scientific data centers have made joint efforts to promote the collection, storage and management, processing and mining as well as opening and sharing of scientific data in different disciplines and areas. In the opinion of **Academician GUO Huadong**, data sharing shall constantly innovate, jump out of the traditional "replication" pattern and build an intelligent data service system integrating data, computing and services, to provide different servicing and data sharing forms for users with different levels of understanding in different areas. In the opinion of **SHI Jiantao, Research Fellow of Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, and Director of Platform for Bioinformation**, besides discoverability, availability and interoperability, the integration of data storage and computing is also necessary. We still have a long way to go in providing users with more calculation convenience in users' shoes.

III. Data Development in a New Phase with Both Opportunities and Challenges

Firstly, the multidisciplinary and trans-regional data interconnection, integration and application is quite promising. According to **Professor Barend Mons**, multidisciplinary data integration could play an important role in driving outputs of higher value and increasing the value of the original data. Based on the global COVID-19 database totally supported by AI, researchers could quickly judge whether it's appropriate to prescribe a patient a certain drug or a new drug on the basis of algorithm analysis without clinical trials. As disclosed by **Director LI Jianhui**, in the next 10 years, with the goal of realizing the intersection, integration, sharing and application of multidisciplinary data, CODATA will establish a trans-regional cooperation network to support future multidisciplinary studies, especially those on significant issues concerning epidemics, climate change, peak carbon dioxide emissions, emission reduction and SDGs.

Secondly, the motivation mechanism for data opening and sharing remains to be improved. Currently, the global scientific community hasn't realized interconnection and resource sharing for lack of an effective mechanism and related technologies. In the opinion of **JIANG Lulu, Product Manager of Science DB, Computer Network**

Information Center of the Chinese Academy of Sciences, currently, the atmosphere of data sharing hasn’t been enlivened among first-line researchers, the research into and practice of data ethics remain insufficient, and massive data opening and sharing as well as international data exchange still face some problems. We shall actively explore the motivation mechanism for data opening and sharing, and intensify the culture of scientific data quotation. According to **HU Lianglin, Deputy Director of Big Data, Computer Network Information Center, Chinese Academy of Sciences, Director of National Public Science Data Center for Basic Sciences, and Secretary-General of CODATA**, as it’s hard for scientific data opening and sharing to step into the phase of “standards first”, we shall call for as much use of the current domestic or international industrial standards as possible to promote the further amplification of the value effect generated by data aggregation. According to **Gergely Sipos, Director of Solution, EGI (European Grid Infrastructure)**, it’s challenging to motivate scientists to share their data with one another. At the moment, the key problem to solve is how to describe data accurately.

Thirdly, infrastructure construction is the key to the success in data opening and sharing. According to **WU Lizong, Associate Research Fellow of National Arctic and Antarctic Data Center**, we lack talents for data center construction, whose cultivation is closely related to powerful infrastructure or research institutes. Infrastructure is far more than hardware, computers, servers or cloud platforms; it involves offering services through the combination of data, standards and hardware. In the opinion of **Director Gergely Sipos**, we need a super platform which combines different computing and cloud methods and integrates scientific applications by the most appropriate approach to meet the complicated and mixed demands for practical application.

Summarized by Zheng Yi