BUILDING A COMMUNITY OF DATA SCIENTISTS: AN EXPLORATIVE ANALYSIS

Lei Liu[1*], Hui Zhang[1], Jianhui Li[1], Runqiang Wang[2], Luqing Yu[2], Jianjun Yu[2], and Peisen Li[3]

*[*1] Secretariat of Chinese National Committee for CODATA, Computer Network Information Center, Chinese Academy of Sciences, 4 South 4th Street, Zhongguancun, Beijing 100190, China*
Email: liulei@codata.cn or liuleirobert369@163.com
[2] Computer Network Information Center, Chinese Academy of Sciences, 4 South 4th Street, Zhongguancun, Beijing 100190, China
[3] China Guanghua Science and Technology Foundation

ABSTRACT

Based on an explorative analysis of the definitions of a data scientist and a community of data scientists as well, this paper points out that with the development of data science and data scientists, the promotion by CODATA and interested partners, the support and efforts of more and more countries and regions, and the drive of social need, building a community of data scientists is full of potentialities and hopes though faced with obstacles and challenges.

Keywords: Data science, Discipline, Data scientists, Community, CODATA

1 INTRODUCTION

Public perception of the new discipline of data science is just in an embryonic stage that is "still in its infancy" (Smith, 2006). Similarly, people rarely mention data scientists or building a community of data scientists. Based on an explorative analysis of the definitions of a data scientist and a community of data scientists, this paper points out that the formation of a community of data scientists depends on the development of data science and data scientists, a development full of potential and hopes. Meanwhile, the process of building such a community is inevitably faced with difficulties and challenges.

2 SOME DEFINITIONS

First of all, let us discuss three questions.

- What is data science? One definition is, "Data science has developed since to include the study of the capture of data, their analysis, metadata, fast retrieval, archiving, exchange, mining to find unexpected knowledge and data relationships, visualization in two and three dimensions including movement, and management. Also included are intellectual property rights and other legal issues." (Smith, 2006) Of course, with the development of the times, data science includes more than the above areas. While according to Journal of Data Science, data science means "almost everything that has something to do with data: Collecting, analyzing, modeling...yet the most important part is its applications--- all sorts of applications".
- What is a data scientist? A tentative definition: a scientist who is devoted to "the study of the capture of data, their analysis, metadata, fast retrieval, archiving, exchange, mining to find unexpected knowledge and data relationships, visualization in two and three dimensions including movement, and management" (Smith, 2006), as well as "intellectual property rights and other legal issues" (Smith, 2006), etc. While according to a research report, data scientists are "people who work where the research is carried out–or, in the case of data centre personnel, in close collaboration with the creators of the data–and may be involved in creative enquiry and analysis, enabling others to work with digital data, and developments in data base technology". (Swan & Brown, 2008)
- What is a community of data scientists? Community is a basic concept in sociology, meaning a group of people of the same religion, race, occupation, etc, or with shared interests. Here community means a group of people of the same occupation. Scientific community/A community of scientists means "a group of scientists having the same faith, values, system objectives, criterions of conduct and spiritual temperament". (Liu, 2006) Accordingly, a community of data scientists is a group of data scientists having the same faith, values, system objectives, criterions of conduct and spiritual temperament. In

this paper a community of data scientists is specially referring to a global one, while communities of data scientists mean communities at various levels. Scientific community/A community of scientists have various forms of social existence: "1) intrinsic forms include scientific school (or school of thought) and 'invisible college'; 2) extrinsic forms include scientific society, scientific research institution, scientific research center, etc.". (Liu, 2006)

3 POTENTIAL AND HOPES

Why is the formation of such a community of data scientists full of potential and hopes? Below are a few major factors.

3.1 The inherent logical result of the development of data science

The formation of a community of data scientists is the inherent logical result of the sound development of data science as a scientific discipline along with the drive of fast development of science and technology (S&T) and the increasingly important strategic value of S&T data, which inspire efforts in constructing the discipline of data science and frequent in-depth scientific exchanges.

3.1.1 The drive of fast development of S&T

Data science, as an academic discipline, is gradually coming into the sight of the academic community though it is not yet well-known. Nowadays, S&T is developing very fast, and tremendous scientific data is being produced everyday worldwide. Accordingly, for the sake of better sharing and use of such data, more and more scientists in various fields besides computer science and informatics are paying more and more attention to data capture, retrieval, mining, archiving, exchanging, application, sharing, database software, etc. Moreover, the rapid development of S&T is increasingly dependent on the development of data science. In short, the disciplines of data science and S&T are rapidly developing by driving each other.

3.1.2 The increasingly important strategic value of S&T data

More significantly, the increasingly important strategic value of scientific data has been recognized. Guo Huadong, Executive Committee member of CODATA, told *China Computerworld*, "Data is a kind of very important resource; it is not only the basis of science and the source of research, but also an embodiment of a country's softpower". (Liu, 2006) Lu Yongxiang, President of the Chinese Academy of Sciences pointed out, "Scientific data has become a kind of important resource and huge social wealth along with mankind stepping into the information age". (Liu, 2006) From basic resources supporting S&T activities, S&T data has become a type of strategic resource supporting the innovative development of S&T. "In international science community of the 20th century, the orientation of scientific data had undergone 4 stages: 1) the stage of start, 2) the stage of unique field of S&T development, 3) the stage of the "pull" of scientific development, and 4) the stage of an independent discipline". In nature such a 4-stage development is "having taken 100 years to realize the strategic status of scientific data". (Sun, 2004)

3.1.3 Inspiring efforts in constructing the discipline of data science

More and more pioneering universities and organizations, etc. are attaching great importance to data science study. Here are only a few representative examples:

Example 1: MIAS Data Sciences Summer Institute - This new computer science summer program, set up by the Department of Computer Science, College of Engineering at the University of Illinois at Urbana-Champaign, aims to prepare future leaders in the data science field. The unique 8-week educational program is designed for undergraduate and graduate students entering the information science field.

Example 2: Datascience @ RPI Computer Science - The CS Data Science Group under the Computer Science Department at RPI (Rensselaer Polytechnic Institute) brings together a diverse set of faculty addressing all aspects of data science from statistical and high-performance issues in data mining to social network modeling and analysis to approximation and randomized algorithms for huge data sets. The growing Data Science Group's outputs are being applied in areas of bioinformatics, astroinformatics, efficient algorithms for price discovery and risk analysis in financial markets, and identification of structure and dynamics in social networks. Moreover, it also supports quite a few PhD students.

Example 3: Data Science Research Group at Osaka University - Focusing on multivariate analysis based on

statistical data and topics of new modeling and statistical causal inference, this research group applies mathematics and computers extensively to study structural equation modeling, graphical modeling, and independent component analysis, as well as survival analysis in biostatistics. Its research includes methodological and application aspects.

Example 4: Data Science Division at Keio University - This data science division is under the Department of Mathematics, Faculty of Science and Technology of Keio University. It is known for its research on statistics in data science and special data science seminar. For instance, it organized an "Australia-Japan Workshop on Data Science 2009" in March 2009, covering a list of important topics.

More and more theoretical exploration on data science is emerging. Here are a few typical examples: (1) Ritei Shibata and his colleagues with Keio University have published ten of a planned Data Science Series of Books since 2001, namely, *Data Literacy*, *Data Sampling*, *Data Mining*, *Model Validation*, *Data Learning Algorithm*, *Spatial Data Modeling*, *Earth Environmental Data*, *Environment and Health Data*, *Clinical Data*, and *Sports Data*, while *Data Modeling*, and *Financial Data* are to appear. In addition, it is worthwhile mentioning that Ritei Shibata studied the relationship among statistics, data engineering, and data science. (2) A paper titled "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics" discussed theoretical foundations and curriculum planning for data science, emphasizing the importance of pedagogy on data science, "A data science department in a university must, of course, concern itself with teaching in its own setting. But it is vital that resources be spent to study pedagogy and to teach pedagogy". (Cleveland, 2001) (3) A commissioned report to the Joint Information Systems Committee (JISC) of UK states, "The skills, role, and career structure of data scientists and curators: An Assessment of Current Practice and Future Needs" (Swan & Brown 2008) studied some vital aspects, that is the definition and career structure of data scientists. This professional assessment report is of great significance.

3.1.4 Frequent and in-depth scientific exchanges

Nowadays, frequent in-depth scientific exchanges can be seen throughout the globalizing world. Science originates from discussion and takes root in exchanges. Likewise, data science came from and develops through academic exchanges promoted by CODATA and multidisciplinary foresighted scientists. In the 21st CODATA Conference more than fifty topics were discussed. In current years, many medium-and small-sized symposiums or workshops such as "the 3rd China-US Roundtable on Scientific Data Cooperation" and "the 1st Australia-Japan Workshop on Data Science" held in March 2009 also covered many hot topics in data science. Academic exchanges associated with data science are deepening with more and more topics.

All kinds of scientific exchanges related with data science can get data scientists to know more and more data scientists, causing relevant discussion and exchanges to improve data science's advancement which will hopefully expand and intensify the circles and social networks of data scientists at different levels, leading to the building of communities of data scientists.

3.2 Persistent promotion from CODATA and interested partners

Since its foundation in 1966, CODATA has been striving to fulfill its mission "to strengthen international science for the benefit of society by promoting improved scientific and technical data management and use". (CODATA, 2006) In the last 42 years, with the cooperation of interested partners, CODATA has organized many activities such as biennial CODATA conferences and various types of meetings and workshops, set up task groups and working groups, and published authoritative S&T data and standards, Newsletters, books, reports, etc.

In 2002, CODATA originated its official publication, the *CODATA Data Science Journal*, which "is a peer-reviewed electronic journal publishing papers on the management of data and databases in Science and Technology". Until now, it has published 8 volumes of papers available free online. "The journal was the most important step taken by CODATA since it was formed" (Smith, 2006), which emphasized the importance of the creation of such a journal leading to a new discipline, data science. Actually the Journal has become more and more influential in the data science field, becoming "a saloon for data scientists and experts in other fields" (Iwata, 2008) and "catalyzing the Data Commons". (Iwata, 2008)

In 2006 more than 600 scientists, engineers, scholars, etc. participated in the 20th International CODATA Conference, the best attended conference in CODATA's 40-year history, which brought about far-reaching consequences to global scientific data sharing and applications and the development of data science.

"CODATA's biennial conference has grown in size and impact, serving as a unique international forum for the advancement of data science across an expanding range of scientific disciplines". (Chen, de Canhos, Guo, Kuznetsov, Makhubela, McMahon, et al., 2007)

Also, CODATA frequently cooperates with other interested organizations such as ICSU, WDC, UN, UNESCO, IAP, TWAS, EU, WSIS, OECD, etc. to promote data and information sharing throughout the world. For instance, CODATA and WSIS regularly cooperate well with each other; the *CODATA Data Science Journal* has acquired financial support from UNESCO; and various kinds of joint conferences or workshops where mutual cooperation between CODATA and other organizations can be seen. As of now, CODATA has 24 National Members, 16 International Scientific Unions, 4 Co-opted Members, and 20 Supporting Organizations. What is more, each National Member has a National Committee for CODATA that includes excellent data scientists and engineers as well as related influential policy makers, etc, organizes and coordinates periodical or unscheduled S&T data activities at national and sometimes bilateral and multilateral levels.

As the most influential international organization that supports the discipline construction of data science, CODATA also attaches importance to fostering a community of data scientists and professionals across the globe, for example a special session of the 21st CODATA Conference titled "Social Aspects: Development of Data Communities and Scientific Capacity Building" that discussed "building a community of data scientists: promise and pitfalls".

3.3 Continuous support and efforts from a variety of countries and regions

The increasingly important strategic value of scientific data has been recognized by more and more countries and regions, institutions and professionals, and continuous support and efforts are being made. For example:
In 2007, the US NSF (National Science Foundation) published a report named *Cyberinfrastructure Vision for 21st Century Discovery*, aimed at "Developing a Coherent Data Cyberinfrastructure in a Complex Global Context" (NSF, 2007), and forwarded its action plan in "The Next Five Years: Towards a National Digital Data Framework" (NSF, 2007), guided by the principles of "A Coherent Organizational Framework - Data Collections and Managing Organizations" (NSF, 2007), "Developing A Flexible Technological Architecture" and "Developing and Implementing Coherent Data Policies ". (NSF, 2007) In 2007, the US NSF also initiated the DataNet Program, for which "up to $100,000,000 over a five year period is expected to be available" (NSF, 2007), to support projects on the sustainable development and application of scientific data.

In the UK in 2004, the Treasury, the DTI, and the DfES published *Science and innovation investment framework 2004-2014*, in response to which the Office of Science and Innovation (OSI) e-Infrastructure Working Group was formed, and in 2007 the Group published a report titled *Developing the UK's e-infrastructure for science and innovation*, aiming to develop national-level e-infrastructure which will "support the entire data and information lifecycle, accelerating the research process itself; support the massive amounts of data both created and re-used by researchers; allow for the re-use and re-purposing of digital data in all their forms". (OSI e-Infrastructure Working Group, 2007)

In the 1980s, the Chinese Academy of Sciences initiated a key project, "Scientific Database and its Application System" (SDB), and after 20 years' development, it became "the most comprehensive scientific information system with largest information quantity, the widest disciplines, and highest service level in the country" (Yu, Xiao & Li ,2006), having made important exploration and innovation in data science and e-Science, with "mass data storage and backup environment and data service platform established, and 503 professional databases established involving 45 institutes, and 21 relevant standards or specifications developed and implemented". (SDB, 2008) In the coming 10-15 years, it has been suggested that the SDB "aim at international forefront, carry out research and practice on data science and data engineering, realize transformation from data to information and knowledge, and lead the development of data science". (SDB Expert Committee, 2006) In 2002, the Ministry of Science and Technology of China initiated the "Scientific Data Sharing Program", a cross-disciplinary, cross-sectoral, cross-regional, cross-industrial large-scale scientific data infrastructure project as an important part of National S&T Infrastructure Platform. The Program aims to set up a national scientific data sharing platform that will "focus on scientific data resources accessed and accumulated by government funding, integrate relevant main databases, construct combined centralized and distributed national scientific data center groups; improve information exchange capacity with international scientific data organizations, promote construction of sharing service network facing diverse innovation entities, and form a hierarchical classified national scientific data sharing service system (MOST, NDRC, MOE, &MOF, 2004)". In 2006 the

State Council issued *the Guidelines on National Medium- and long-term Program for Science and Technology Development (2006-2020)*, which sets the proportion of research and development expenditures at 2.5 percent of the GDP. The country will continue to strengthen S&T Infrastructure Platform construction, among which one of the emphases is to "construct a scientific data and information platform" (The State Council of PRC, 2006) and set up effective sharing mechanisms of the S&T Infrastructure Platform.

In short, the increasing efforts and support for scientific databases or data infrastructure by more and more countries have fostered a large number of data science professionals and will continuously require more and more people to engage in data-oriented management, applications, research, etc. For instance, "about 500 master-degree and 300 doctoral postgraduates have been trained only through the SDB Project in the past 20 years". (SDB Expert Committee, 2006) There is no doubt that all this is creating important opportunities for building communities of data scientists at different levels.

3.4 Drive of social need

Social need is one major driving power that accelerates the formation and development of scientific communities. "Social progress and needs is major power that accelerates the formation and development of scientific communities". (Han, 2009) According to relevant research by CAST (China Association for Science and Technology), the earliest association in the world is the Royal College of Surgeons in Edinburgh founded in Scotland in 1505, while the earliest association in China is Yi Ti Tang Zhai Ren Yi Hui founded in Beijing in 1568. Coincidently, both of the two earliest associations belong to the medical field, which reflects that medicine is very closely connected with human immediate interest, and social need causes medical associations to come into being.

Similarly in society, all kinds of increasingly complex data-related projects and undertakings urgently need the illumination of data science as a beacon, especially directions from communities of data scientists, the shaping of which is conducive to spread and develop data science and technology to benefit the society, enhancing efficient applications of data resources to accelerate scientific and technological innovation and development. Actually, more and more communities of data scientists aimed at a certain social need have come into being. For instance, The MIAS Data Sciences Summer Institute is funded by the US Department of Homeland Security (DHS), in hopes of finding solutions that aid the DHS mission; the CAS Research Center on Fictitious Economy & Data Science focuses on seeking the theory of data science from fictitious economy, knowledge economy and regional economy phenomena, etc. and has completed many projects about applications on banking, insurance, financial risk management, etc., which meet social needs at home and abroad.

In a word, the sound development of the young discipline of data science, the promotion from CODATA and its partners, the insistent efforts of so many countries and regions, as well as the strong drive of social need will undoubtedly enhance the growth of data scientists and the shaping of a community of data scientists.

4 DIFFICULTIES AND CHALLENGES

Inevitably, building such a community is faced with difficulties and challenges:

4.1 The time-consuming developing process of data science

We can find many more results of "scientific data" than "data science" using a search engine like Google, etc. Meanwhile, we can find very few results for "data scientist". Though there are many data centers or databases worldwide and data science research groups/institutes/divisions are emerging, as mentioned in 3.1.3 of this paper, we still have comparatively fewer research institutions called "data science institute/research center" or something similar. We can find very few academic journals with data science in their titles except for the CODATA *Data Science Journal* and the *Journal of Data Science*. We have too few formal courses or textbooks with the words data science in universities or institutes. Actually, "to be taken seriously, any discipline needs to have endured over time". (Smith, 2006) "The term 'Computer Science' also evolved after a long time and only dominated in the early 1970's". (Smith, 2006) Therefore, similar to that of computer science, the development of data science as well as its public perception and popularization, fostering of students, professionals, and scientists, etc. need a long time to come to fruition.

4.2 Comparatively fewer data scientists

A community of data scientists is based on the appearance of professional data scientists, but comparatively speaking, there have been very few data scientists up to now. There are many scientists in various areas such as physics, chemistry, computer science, etc. but many fewer professional data scientists. Though many people are devoted to scientific data and regard themselves to be scientific data workers or experts, few of them consciously claim to be a data scientist. Why? There is not enough academic acceptance of data science as an entity, especially lacking are related necessary theoretic exploration and wide debate. Common consensus on the title, definition or career of a data scientist is absent. “There is no defined career structure for data scientists and this is a major problem that must be resolved …” (Swan & Brown, 2008)

Because data science cuts across many disciplines and faces the challenge of requiring interdisciplinary talents, scientific data workers and experts in different fields need to readily transform themselves into professional data scientists all the while remaining true to their physicist, chemist, computer scientist, etc. origins.

4.3 Challenges for CODATA

It should be noted that CODATA still has a long way with its membership from the present 24 National Members and it needs to attract more people from more countries to participate in CODATA activities so as to help foster more communities of data scientists at the national level and to build a community of data scientists at the global level. The CODATA *Data Science Journal* can be expected to play an even more important role to promote the data science community.

4.4 Challenges for community construction

As mentioned in Part 2, the scientific community has various forms of social existence including scientific schools, “invisible colleges”, scientific societies, scientific research institutions, scientific research centers, etc. Actually, the quantity of “invisible colleges” related to data science is difficult to calculate. There are a great number of scientific data centers or databases, etc. in the world but no scientific school expressly for data science, no formal scientific society named “data science society”, and not enough data science research institutions and centers as mentioned in Part 3.1. Perhaps an international data science society under CODATA’s guidance, which naturally includes the existing members and partners of CODATA, and more importantly, admits even more individual scientists across the world, is one choice for the resulting global community of data scientists.

5 CONCLUSION

In the information society, scientific data and knowledge are becoming increasingly important, attracting efforts and support from more and more countries, organizations, institutions, and individuals. Despite difficulties and challenges, the discipline of data science is becoming widely recognized, and more scientists in various disciplines or fields are engaging in the field of data science and becoming data scientists in the true meaning of the words. Hopefully, a global community of data scientists will come into being to benefit the world.

6 ACKNOWLEDGEMENTS

Sincere thanks must be extended to Prof. Shuichi Iwata with University of Tokyo who greatly enlightened the author to make necessary revisions to this paper.

7 REFERENCES

Chen, R., de Canhos, V., Guo, H., Kuznetsov, F., Makhubela, L., McMahon, B., et al. (2007) *CODATA Strategic Plan 2006-2012.* Retrieved April 3, 2009 from the World Wide Web: http://www.codata.org/Strategic-Plan-Public-Review.pdf

Cleveland, W. (2001) Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *ISI Review* 69:21-26.

CODATA (2006) Mission Statement approved at the 25th CODATA General Assembly. Retrieved March 5, 2008 from the World Wide Web: http://www.codata.org.

Han, Qide (2009) Promoting the Development of Chinese Scientific and Technical Societies Based on the Regular Patters of Scientific Community. Science and Technology Review, Foreword.

Iwata, S. (2008) Editor's Note: Scientific "Agenda" of Data Science. *Data Science Journal Vol.7*, 54-56. Retrieved April 3, 2009 from the World Wide Web: http://www.jstage.jst.go.jp/article/dsj/7/0/54/_pdf

Liu, Jin'ge (2006) Chapter 3: The Main Bodies and Social Support of Scientific and Technological Activities. In Zhu, Shaojun (Eds.) *Scientific and Technological Dialectics and Methodology*. Beijing, China: People's Military Medical Press.

Liu, Jingjing (2006). Scientific and technical data sharing tests national softpower. *China Computerworld. Oct.30 (A25)* Retrieved March 5, 2008 from the World Wide Web: http://www2.ccw.com.cn/06/0642/a/0642a15_7.html

NSF (2007) *Cyberinfrastructure Vision for 21st Century Discovery*. Retrieved 3 April, 2009 from the World Wide Web: http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf

NSF (2007) Sustainable Digital Data Preservation and Access Network Partners (DataNet). Retrieved April 5, 2009 from the World Wide Web: http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm

SDB (2008) *Introduction of Scientific Data Center*. Retrieved 16 April, 2009 from the World Wide Web: http://www.csdb.cn/view.jsp?tipid=1126059532207

SDB Expert Committee (2006) The Scientific Database of Chinese Academy of Sciences: Twenty Years Review. In Scientific Database Project Office (Eds.), *Proceedings of Scientific Database and Information Technology. Vol.8.*, Beijing, China: China Environmental Sciences Press

Smith, F. (2006) Data science as an academic discipline. *Data Science Journal Vol.5*, 163-164. Retrieved March 5, 2008 from the World Wide Web: http://www.jstage.jst.go.jp/article/dsj/5/0/163/_pdf

Sun, Honglie (2004) *Recognizing the strategic status of scientific data*. Retrieved Oct. 2, 2008 from the World Wide Web: http://www.gmw.cn/01gmrb/2004-03/12/content_5184.htm

Swan, A. & Brown, S. (2008) *The skills, role and career structure of data scientists and curators: An Assessment of Current Practice and Future Needs.* Retrieved March 5, 2008 from the World Wide Web: http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf

The Ministry of Science and Technology (MOST), National Development and Reform Commission (NDRC), the Ministry of Education (MOE), and the Ministry of Finance (MOF) (2004) *National Science and Technology Infrastructure Construction Guidelines (2004-2010*). Retrieved April 16, 2009 from the World Wide Web: http://news.xinhuanet.com/st/2004-09/25/content_2019790.htm

The Office of Science and Innovation (OSI) e-Infrastructure Working Group (2007) *Developing the UK's e-infrastructure for science and innovation*. Retrieved April 3, 2009 from the World Wide Web: http://www.nesc.ac.uk/documents/OSI/index.html

The State Council of PRC (2006) *the Guidelines on National Medium- and long-term Program for Science and Technology Development (2006-2020).* Retrieved April 5, 2009 from the World Wide Web: http://www.gov.cn/jrzg/2006-02/09/content_183787_9.htm

Yu, Luqing, Xiao, Yun, & Li, Jianhui (2006) Research and Implementation of SDB Data Sharing Policy. In Scientific Database Project Office (Eds.), *Proceedings of Scientific Database and Information Technology Vol.8.*

Beijing, China: China Environmental Sciences Press.

(Article history: Received 12 January 2009, Accepted 24 April 2009, Available online 30 September 2009)