Study on the University Data Literacy Education under the E-Science Environment

Wei DU\textsuperscript{a} Qiang LIU\textsuperscript{b,}\text superscript{*} and Xiao-Ying MA\textsuperscript{c}

University of Jinan Library, Jinan, Shandong, China

\textsuperscript{a}dw_163@163.com, \textsuperscript{b}lib_liuq@ujn.edu.cn, \textsuperscript{c}lib_maxy@ujn.edu.cn

\textsuperscript{*}Corresponding author

Keywords: E-Science, University, Data literacy, Education.

Abstract. Under the e-Science environment, the importance of scientific data has become increasingly prominent. The characteristics of e-Science determine that researchers must possess high data literacy. The need of e-Science environment for data literacy is discussed from the connotation and cause of data literacy, and how to carry out data literacy education under e-Science environment from the educational content and educational forms is also discussed.

Introduction

Under the e-Science environment, with the rapidly growth of the scientific data, a huge amount of data have produced, which have been deeply analyzed and excavated. As a result, a new research paradigm—data-intensive research—has formed. Under such an environment, researchers face more and more problems with data management, data publishing and data reference and so on. The acquisition, use and evaluation of data have become the basic qualities that researchers should possess. Data literacy has gradually become an indispensable factor and a core competitive force for enhancing the capability of scientific research innovation.

Overview of E-Science

The interaction between information technology and scientists leads to e-Science, in which researchers use many different methods to collect and produce data, whose distributed and large-scale global cooperative scientific research ways will help researchers access a large amount of data, massive computing resources, and high-performance visual environments and so on. Through e-Science technology, researchers can effectively access and manage geographically distributed data and network resources, and analyze these massive data, such as data mining and knowledge discovery in time. In the specific big data environment of e-Science, the huge data and the complicated data management problems make researchers face severe challenges. The issues such as data production and collection, data analysis and processing, data saving and archiving, data publication and sharing are throughout the whole process. Researchers need data literacy training and guidance urgently and they are required to have the appropriate data awareness and macro control further.

The Needs of E-Science for Data Literacy

The Connotation of Data Literacy

Before introduce the information literacy, we introduce the research data firstly. Research data are generated for different purposes and through a range of processes. It can include but is not limited to: ① Experimental or laboratory data; ② Simulations generated from test models; ③ Derivations or compilations e.g. text and data mining, databases, 3D models; ④ Collections of peer reviewed datasets published e.g. chemical structures, gene sequence databanks, or literary research; ⑤ Published research results including tables underpinning the published work[1]. Data literacy is
also known as "quantitative literacy", "data literacy", "statistical literacy", "scientific literacy" and so on, whose definition has not gotten a unified consensus. According to the research and development of literature both at home and abroad, data literacy is the awareness and ability to discover and acquire data properly, select and evaluate data critically, regulate and manage data normatively, use and share data reasonably[2]. It is important to emphasize that data literacy encompasses not only data capabilities but also data awareness and data ethics. That is to help people to access and reflect on various sources and types of data, and systematically make use of data content and expertise to gain knowledge, to facilitate the transfer from data to specific academic research goals or research findings, and to help people develop data sharing and academic community structures that support knowledge exchange.

The Reason of Data Literacy Rise

Scientific Research Evolves toward a Data-Intensive Paradigm. Since 2000, with the exponential growth of scientific research data in the e-Science environment, the value and power highlighted by the data-driven innovation and decision-making has been gaining a high attention in various fields such as economics, management and scientific research. Based on e-Science, the fourth paradigm: data-intensive scientific discovery states that scientific research has entered the fourth paradigm, that is the fourth paradigm of scientific research based on data-intensive computing. Data-intensive science consists of three basic activities, such as collection, management and analysis. In this paradigm, with new scientific research methods researchers will accelerate scientific discovery and accelerate more natural interaction between data consumers and the data[3]. In such a data environment, the capability of data acquisition, analysis, management and sharing and the rational use of data become the necessary essential literacy of scientific researchers.

The Need of Scientific Research Data Management. With the rapid development of science and technology, all disciplines have the tendency of producing a large amount of primary data and derived data, and at the same time, the complex and diverse of data types have become inevitable for the development of scientific research under the e-Science environment. At present, there is still a problem that a huge amount of data are lost. Therefore, in order to avoid the losses of massive data, it is necessary to strengthen the data warehousing, storage and sharing, and establish a long-term and sustainable knowledge base of scientific research data. The sound data management procedures can: 1. Increase the efficiency of research; 2. Help guarantee the quality and authenticity of data; 3. Enable the exposure of research outcomes through collaboration and dissemination; 4. Provide for the reproducibility of experimental and computational outcomes; 5. Facilitate the validation and verification of results[1]. With the increasingly obvious impact of data management, not only the demand for data management professionals continues to increase, but also the higher requirement for the collaboration between researchers and data managers is put forward. Researchers must have some data management skills, therefore, the data literacy education of researchers is imperative.

The Data Policy Formulation of Government Funding Agencies. Many research fields lack established norms and practices in managing, sharing and preserving their data, making it difficult for researchers to respond to these pressures effectively. In some countries in Europe and the United States, the government funding agencies set forth the requirements and norms for the management of scientific research data at the national level from the perspective of resource sharing and interdisciplinary cooperation. For example, in 2007, the Office of Science and Innovation (OSI) in UK established an e-infrastructure group with wide participation including from higher education, the group published its report Developing the UK’s e-infrastructure for science and innovation, which clearly indicated the areas requiring further development, including data curation and management. The RUGIT and RLUK feasibility study aimed to build on the OSI report by addressing the issue of the entire data management lifecycle and providing a roadmap for infrastructure development[4]. Many research funders in some countries are putting policies in place to ensure that datasets judged to be potentially useful to others are curated in ways that are allow discovery, access and reuse. The development of data policy undoubtedly raises higher
demands on the work and capabilities of scientific researchers in managing and processing scientific research data.

**The Urgent Requirement for the Publication of Academic Achievements.** Scientific data sharing around the world is imperative. To achieve the data sharing, the publication of scientific data is also imperative. Data publishing refers to the peer-reviewing research data from a scientific perspective and publishing it publicly, creating standard and the permanent data citation information for reuse and citation by other researchers[5]. The current data publishing can be viewed from two aspects. On one hand, from the publication of research papers, some periodicals require authors to have the original data of the papers when submitting them. On the other hand, from the output of scientific research, scientific research data are an important form of scientific research. In the new academic exchange environment, the data journals become an important carrier of scientific research data, such as the ScientificData of Nature Publishing Group, the GenomicsData of Elsevier. This shows that academic achievements put forward higher requirements for the researchers’ data literacy.

**Data Literacy Education under the E-Science Environment**

**The Education Contents**

Specifically, the data literacy education should have the following aspects. The first is the basic education which includes: ① The importance of data. ② The data lifecycle. ③ The Planning for data management. ④ Documenting your data. ⑤ Analyzing your data. ⑥ Managing secure and private data. ⑦ Short-term storage. ⑧ Preserving and archiving your data. ⑨ Sharing/publishing your data. ⑩ Collaborations and data. ⑪ Reusing data[6]. The second is the data standardization education based on the basic data-knowledge, including the universal education on user's data collection methods, collection specifications, processing specifications, data types, data formats, data association and other related basic knowledge, which can provide the necessary help for researchers to understand and standardize the data. This can make the education of the resulting research data such as data formats, generic identifiers, simplified standards, naming conventions, and other data organization education more standardized and easy to organize.

**The Education Methods**

**Open the Data Literacy Courses.** (1) Classroom teaching. ① Some foreign universities have begun to systematically open the data literacy course. For example, data management courses such as the Data Management Course of University of Minnesota and the Research Data Management Program of New England provide targeted data literacy education for different types of scientific researchers and undergraduates. ② Some foreign universities integrate data literacy into other classes. For example, Maybee Clarence, et al. found that the teachers of nutrition science and political science conducted data literacy education in different levels for students[7]. ③ Data literacy education courses can be conducted in form of open class, such as to invite experts of various disciplines to give lessons to improve the quality of scientific data literacy education in order to enhance the whole staff’s awareness of data sharing, and provide data ability to collection, collation, retrieval, reuse in the scientific research process.

(2) Online class. Online class in form of MOOC and micro-class is one of the main development direction of higher education. Many European and American universities have opened MOOC to offer data science, data science management, data analysis tools, machine learning and data mining courses[8]. For example, University of Minnesota Libraries offer free drop-in workshops on data management best practices periodically throughout the year. The contents of their Online Data Management Course include: ① Introduction to data management; ② How to inventory, store, and backup your data; ③ How to create data that you (and others) can understand; ④ How to navigate rights and ownership of your research data; ⑤ How to share your data and ethically reuse data created by others; ⑥ How to digitally preserve your data for the future; ⑦ Complete your DMP (Data Management Plan) [9].
Establish the Data Management Navigations. Data management navigations are to facilitate users to quickly and correctly identify and obtain the reasonable data management resources under certain guidelines, and further provide a reasonable and effective data services, which are not only the main channel for researchers or students to learn about data management-related knowledge, but also an important complement to the data literacy education and teaching resources. Researchers can organize and understand the concept and process of data management, and also establish a comprehensive awareness of data management through their thematic information organization. Professional teachers or librarians can also use them as a flip teaching, problem-based learning, task-based learning and other teaching organizational model of the pre-class or after-school reading materials. Some university libraries in the United States, Australia and the United Kingdom use the libguides to establish the scientific data management navigation. For example, MIT Libraries create a Data Management Guide helping MIT faculty and researchers manage, store, and share data they produce[10].

Carry out Advisory Services. The data experts can provide SDI services about data management, and the users can obtain the support and help in the research data management, data management policies, data management plans and so on by way of booking, visiting or participating in the research team. For example, the Sheridan Libraries of Johns Hopkins University set up Data Management Consultant and Data Services Manager to specialize in data management services[11]. Data services at University of Virginia Library are evolving to meet emerging research needs. The experts at University of Virginia Library offer workshops and other forms of training, and are available for one-on-one consultation[12].

Hold Seminars about Data Literacy. The seminar is a comprehensive teaching method that focuses on the collection of materials, scene design, demonstration, case analysis and group discussion around a certain topic, which aims at cultivating participants' autonomous learning, practice and teamwork. Through the seminar, people with similar interests can be organized to discuss the content and training methods of data literacy education in order to promote the data literacy education better and raise the data awareness of scientific researchers. The "Data Literacy Seminar" organized by Purdue University, Cornell University library, University of Minnesota and University of Oregon is particularly typical.

Conduct a Lecture. According to the needs of different users, different lectures related to data literacy can be carried out. For science and engineering users, we should focus on teaching data awareness, data processing, and software applications. For humanities and social science researchers, we need to focus on the training of their data awareness, data acquisition, and outcomes expression and so on. For example, UQ Library of University of Queensland can give advice, training and options on how to: ① Create and manage an online data management plan; ② Meet funder and publisher requirements; ③ Demonstrate research integrity; ④ Store and archive your research data; ⑤ Create a record and set access rules for your data to make it discoverable[1].

Conclusion

Data literacy is the necessary awareness and ability requirements for the scientific and technical researchers under the e-Science environment. However, because of the different social environment in different countries, the emphasis on data literacy is also different. There are big differences in the formulation of relevant policies and norms, the development of education and the emphasis of scientific researchers. Many colleges and universities in some foreign countries have carried out the course of data literacy education. Although some colleges and universities in China have conducted a lot of research on data literacy education, its real implementation is negligible. On one hand, perhaps it is because the leader does not pay much more attention to it or there are not many sources of funds. On the other hand, it is because there are still not enough professionals who can really do data literacy education or training. So, firstly, China need to increase policy, funding and personnel support for data literacy education, because the Country has the important responsibility to promote the policy development of data access and reuse at all levels and to facilitate the flow of information. Secondly, there is a need to strengthen the data literacy training for the entire industry.
practitioners or the librarians who mainly engage in this work in order to adapt to the needs of conducting scientific researches under the e-Science environment.

References


