Data Literacy Education Design Based on Needs of Graduate Students in University of Chinese Academy of Sciences

Wu Ming[1,2](✉) and Hu Hui[1]

[1] National Science Library, Chinese Academy of Sciences, Beijing, China
{wum,huhui}@mail.las.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China

Abstract. In the new data-intensive research environment, research data is an important part of scientific findings and every researcher will face sophisticated data management problems during their research life. Solving these data problems requires researchers and students have new skill sets and competencies, which ensure their outputs are accessible, discoverable and reusable. Using the online questionnaire survey method, we conducted a data literacy survey among 59 graduate students of life science in University of Chinese Academy of Sciences (UCAS). The current situation and needs of graduate students' data literacy competences are revealed. On the basis of demand investigation, the data literacy education model of teachers, students and curriculum is constructed, the education content is based on research data lifecycle and includes three levels of learning modes. In addition, the data literacy education implementation scenes for graduate students in UCAS were also designed, and provide evidence for libraries to implement data literacy education services better.

Keywords: Data literacy · Research data management · Data literacy education
Graduate students

1 Introduction

In the new data-intensive research environment, research data is an important part of scientific findings, and every researcher will face sophisticated data management problems during their research life, such as data generation and collection, data documentation and processing, data storage and backup, data publishing and sharing. Data literacy has become new skill sets and necessary competencies for researchers and students to solve these data problems [1]. At the same time, many funding agencies, universities and other organizations now have a research data management policy in place, such as the NSF Data Management Plan Requirements which states, "Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled 'Data Management Plan'. This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results." And in NIH's view, all data should be considered for data sharing. Data should be made as widely and freely available as possible while

S. Kurbanoğlu et al. (Eds.): ECIL 2017, CCIS 810, pp. 158–168, 2018.
https://doi.org/10.1007/978-3-319-74334-9_17

safeguarding the privacy of participants, and protecting confidential and proprietary data. To facilitate data sharing, investigators submitting a research application requesting $500,000 or more of direct costs in any single year to NIH on or after October 1, 2003 are expected to include a plan for sharing final research data for research purposes, or state why data sharing is not possible. Requirements by different funders to make research data available vary by country, institution, or discipline.

Graduate students are a natural audience for educational programming on data literacy education issues. In the STEM disciplines, graduate students are often expected to carry out most or all of the data management tasks for their own research, and frequently participate in data activities to support team projects [2]. That is to say, this requirement demonstrates that data management skills are needed in a wide range of disciplines and that core skills, as well as discipline-specific training, should be embedded into the graduate curricula [3]. Graduate students are being required to improve and enhance their research data management skills and practices, and data literacy services offer librarians an opportunity to expand their role in the research enterprise within an institution.

In recent years, there has been growing discussion in the literature about how to develop data services and data literacy education in academic libraries. We identified several best practices for teaching data literacy in the literature. The New England Collaborative Data Management Curriculum (NECDMC) is one example of a data management curriculum [4]. Designed collaboratively by librarians from New England academic institutions, it was created for the purpose of using one or more modules to instruct other librarians and researchers about the need for, and use of, good research data management practices. The project, named data information literacy (DIL), consisting of research teams from Purdue University, the University of Minnesota, the University of Oregon, and Cornell University, aims to develop and implement a DIL curriculum in conjunction with university faculty to address these needs [5]. The three central goals for this project are to build infrastructure in the library community for DIL skills, to have students learn DIL skills appropriate to their disciplinary context, and to develop a robust process for librarians to articulate DIL curricula in their research communities. Based on research conducted at Purdue, DIL seeks to incorporate and build upon relevant aspects of information and other literacies to articulate the skill sets needed by graduate students to fulfill their obligations and engage their communities of practice. A central tenant of DIL is the recognition of researchers as producers of data, as well as data consumers. The DataTrain project at the University of Cambridge, aims to build on findings and tools developed in the Incremental project (JISC 07/09 funding strand) by developing disciplinary focused data management training modules for post-graduate courses in Archaeology and Social Anthropology at the University of Cambridge [6]. Another well-known curriculum called MANTRA is an online course from the University of Edinburgh. MANTRA is a free, online non-assessed course with guidelines to help researchers understand and reflect on how to manage the digital data they collect throughout their research. It has been crafted for the use of post-graduate students, early career researchers, and also information professionals [7]. It is freely available on the web for anyone to explore on their own. There are eight online units in this course and one set of offline data handling tutorials. Each unit takes up to one hour, plus time for further reading and carrying out the data handling exercises.

Many libraries have taken on the role of providing instruction in data literacy, but few libraries have addressed the practice aimed at providing data literacy instruction at the life science field. In addition, many libraries have directed their training efforts toward students at the undergraduate and graduate level, rather than focusing on the students' data management behavior and needs in order to design the data literacy instruction. So, this paper chose the graduate students in UCAS as the research objects, designed a questionnaire to investigate their data management behavior and data literacy needs, designs data literacy education implementation scenes and a service model for graduate students in UCAS, and provides recommendations for libraries to improve their implementation of data literacy education services.

2 Methodology

2.1 Questionnaire

We conducted an online survey using Sojump that consists of 30 questions regarding graduate students' own basic information, data management behavior, attitudes, and education related to managing research data. The 12 competencies of Data Information Literacy [2] are used as a guide in designing the survey; further consideration is given to research data lifestyle and the specific situation of graduate students in UCAS. The questionnaire includes two sections and 30 questions. Section one is about personal information of the respondents (Q1–Q3). Section two involves eight parts:

(1) research data basic knowledge (Q6–Q9)
(2) data management plan (Q10–Q13)
(3) data collection and documentation (Q14)
(4) data processing and analysis (Q15–Q18)
(5) data management and preservation (Q19)
(6) data publishing and sharing (Q20–Q21)
(7) research data ethics (Q22–Q23)
(8) needs for data literacy course (Q24–Q30).

In the questionnaire, 28 of the questions are closed-ended in order to facilitate responding and analysis, with options for 'Other' presented where applicable. Two of the questions are open-ended text boxes to obtain additional descriptive information from the participants.

2.2 Participants

We conducted a data literacy survey among 59 graduate students of life science in UCAS. The selection of the respondents mainly takes into account two aspects. The first aspect is different data management behaviors among different disciplines. Life science is one of the main disciplines in Chinese Academy of Sciences and recruits thousands of graduate students including the School of Life Sciences of UCAS, as well as Institute of Botany, Institute of Zoology, Institute of Genetics and Development, Institute of Microbiology and others. Furthermore, in the field of life sciences, the process of research data creation and collection is complex. The analysis and processing of research data is also complicated, and researchers and graduate students face huge amounts of research data and complex data management issues. So this survey chooses life sciences as a starting point, in order to understand their data management behavior and needs. The second consideration is the necessity and feasibility for graduate students in UCAS to receive data literacy instruction. In UCAS, graduate students should finish all of their courses, both compulsory and elective, in their first graduate year, but have not yet started their research work at this stage. Therefore, preparing for future scientific research work, it is necessary and feasible first to cultivate students' research data management ability through credit courses and training lectures.

Our participants come from 22 research institutes in the field of life science of the Chinese Academy of Sciences, such as the Institute of Botany, Institute of Genetics and Developmental Biology, Chengdu Institute of Biology, South China Botanical Garden, Institute of Zoology and others. The research area of participants involves Genetics, Zoology, Biochemistry and Molecular Biology, Botany, Bioengineering, Ecology, Developmental Biology, Cell Biology, Genomics, Marine Biology, and Biochemistry. Overall, the selection of survey samples is in line with the CAS in the field of life science research distribution and has a good representation.

3 Results

3.1 The Status of Graduate Students' Data Literacy

This part intends to offer an overview of the status of graduate students' data literacy, mainly reflecting the data management behavior of graduate students at different stages of the research data lifecycle, including research data basic knowledge, data management plan, data collection and documentation, data processing and analysis, data management and preservation, data publishing and sharing, and research data ethics.

Basic Ideas About Research Data

The Importance of Research Data Management. The question setting and answer option were referenced from the UK Data Archive. Through the investigation of graduate students on the significance of research data management, we can reveal students' awareness and positive attitude on research data management. The survey results indicate that most of the graduate students in the field of life science have a full awareness of the importance of research data management. Among them, the graduate students believe that the main significance of managing their data is promoting innovation and potential (89.83%), reducing the cost of duplicating data collection (86.44%), and encouraging improvement and validation of research methods (83.05%).

Knowledge of Discipline Research Data. This question examines whether students understand and how much they know about research data in their discipline. According to the survey results, most of the graduate students can determine their discipline research data, and statistical and measurement data (88.14%), experimental and simulation results (86.44%) comprise the main data in the life science field.

Policies and Requirements of Research Data Management and Sharing. This question examines whether students understand and how much they know about data management policy and requirements. The survey results indicate that some of the graduate students in the field of life science say they know data management policies and requirements (74.58%), include data policies and requirements in their laboratory (50.85%), and research organization (33.90%). Overall, graduate students lack knowledge about data management policies and requirements, especially regarding journal publishers, data repositories, research funding agencies and other data management policies and requirements.

Research Data Lifecycle. Data often have a longer lifespan than the research project that creates them. Researchers may continue to work on data after funding has ceased. Follow-up projects may analyze or add to the data, and data may be re-used by other researchers. Well organized, well documented, preserved and shared data are invaluable to advance scientific inquiry and to increase opportunities for learning and innovation. So, it is necessary for graduate students to know the concept of the research data lifecycle. According to the UK Data Archive, research data lifecycle includes creating data, processing data, analyzing data, preserving data, giving access to data and re-using data. This question mainly investigates whether the they understand the concept of the research data lifecycle. The survey results indicate that only 20.34% graduate students know something about what the research data lifecycle is, but most of them feel familiar with the research data lifecycle.

Data Management Plan. A data management plan is a formal document that outlines what you will do with your data during and after you complete your research. It describes the data that will be created, the standards used to describe the data (metadata), who owns the data, who can access the data, how long the data will be preserved (and/or made accessible), and what facilities and equipment will be necessary to disseminate, share, and/or preserve the data (NCSU Libraries).

Components of Data Management Plan. The question setting and answer option referenced from NSF, and a data management plan may include:

- the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project
- the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies)
- policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements
- policies and provisions for re-use, re-distribution, and the production of derivatives

- plans for archiving data, samples, and other research products, and for preservation of access to them.

This question mainly examines whether students understand the concept and specific content of the data management plan. According to the survey results, 90% of the graduate students understand what a data management plan is and can determine the content contained in the data management plan, but there are still 10.17% of students who have no idea about data management plans.

Data Collection and Documentation

The Methods of Recording and Storing Research Data. This topic is mainly to investigate the data collection methods in life science, as well as how the graduate students collect and document their research data during research practice. The results of the survey indicate that, almost all of them have mastered how to collect and record research data. Among them, the most common means of data recording are USB flash disk or mobile hard disk and other portable storage devices (94.92%), personal computer (94.92%), paper or laboratory notebooks (93.22%), and laboratory or office computer (86.44%).

Data Processing and Analysis

Tools and Software of Processing Research Data. The types and formats of life science research data are various and so are the tools and software for processing data. The results of the survey show that data processing and analysis in life science uses basic data processing tools such as Excel, SPSS, and R, and also some special tools for life science such as Primer, Origin Demo, SigmaPlot, STATISTICA, Clustalw, Curve Expert, Click It Graph, Graph Pad, and PRISM. It should be noted that there are still 10.17% of the graduate students who do not know or are uncertain of which tools or software are often used in processing their research data.

The Sources of Research Data. Life science is a complex discipline with large amounts of data and diverse data sources. This topic focuses on investigating graduate students' data sources in their research practice, the results of the survey, Some graduate students have a wide range of research data sources, both their own and team experimental data, but also datasets downloaded from literature, data centers, and data warehouses (59.32%). But we need to pay attention those students who just create their research data personally or rely on their research team (37.29%).

The Standards to Evaluate the Quality of Research Data. Quality control is an important part of research data management. An examination of students' understanding of methods of controlling and evaluating the quality of research data, can reflect their ability level at managing research data. The survey results indicate that the vast majority of graduate students have mastered the methods of data quality control and evaluation, including data authenticity, data integrity, data normative, data reproducibility.

Data Management and Preservation

The Description of Research Data (Metadata). Using metadata to describe research data will keep it understandable and reusable, So, this topic mainly examines attitudes and status of graduate students at understanding and using metadata for description of research data. According to the survey results, most students can describe research data normatively in research practice. They always describe their data following the rules in their research team or laboratory (94.92%), or create data description rules by themselves. However, there are still some students who have never described their research data.

Data Publishing and Sharing

The Way to Publish Research Data. It has become a trend for researchers to submit data in the format the academic journals request when publishing papers. As the survey results indicate, most of the graduate students know how to publish their research data, including submitting it as supporting information for the paper and providing the DOI (76.27%), submitting to the data repository or institutional repository designated (52.54%), and publishing data papers in data journals (50.85%). But on the other hand, there are still some students who say they do not know or are uncertain of how to publish research data.

Research Data Ethics

The Behavior of Data Reference. When we quote research data from others, we should reference the data source in the same manner as citing journal papers. This question mainly investigates graduate students' awareness and behavior when referencing data from someone else. As the survey results indicate, most of them would cite data source normatively (96.61%), and acknowledge the authors of the data sources (52.54%). But there are some students who say they never indicate the data source or do not know how to cite research data normatively.

Research Data Ethics. Many data ethical issues are involved during managing and sharing research data. This topic mainly investigates whether graduate students understand the ethical issues of research data. According to the survey results, many of them know about research data ethics, such as research data ownership (79.66%) and right to informed consent (76.27%), but just a small number of students know about commercial interests, security secrets, and privacy. At the same time, there are 13.56% of students who show that they know nothing about research data ethics.

3.2 The Needs of Graduate Students' Data Literacy

We have investigated graduate students' status of data literacy in life science around the research data life cycle, and now we have several question about what they want to know about research data management, in what way to offer data management courses, and suggestions for data literacy courses.

Attitudes Towards Data Literacy Education. The answer to this question is no doubt yes (essential, 81.36%, important, 18.64%). All graduate students think it necessary to have data literacy education before starting their research and hope to improve their research data management skills.

The Contents Data Literacy Education Needs. We will design course syllabi based on the research data life cycle, so we investigate the specific learning needs through different data management phases. The results of the survey (see Table 1):

Table 1. The contents data literacy education needs

Data management phase	The contents data literacy education needs	%
Data management plan	Elements of data management plan	61.02
	Tools of data management plan	89.83
	Policies and requirements about data management and sharing	81.36
	Research data life cycle	71.19
Data collection and documentation	The type, format and data volume of research data	94.92
	Quality control and evaluation of research data	83.05
	How to document and describe research data in fields	81.36
Data processing and analysis	Tools to process and analyze research data	98.31
	Data visualization	76.27
Data management and preservation	Research data security	91.53
	Research data store and backup	88.14
	Metadata standards for describing research data in fields	81.36
	Research data naming rules	62.71
	Research data version control	47.46
Data sharing and reuse	Retrieval and acquisition external research data	98.31
	ways to publish and share research data	88.14
	Standard for citation of research data	79.66
	Ethics related to research data and protection measures	77.97
	Research data sharing and license agreement	69.49
Teaching approach	Lecture	69.49
	RSS/Alerts	67.80
	Online course	66.10
	Credit course	61.02
	WeChat public, blog and other media	50.85
	Workshop	27.12

4 Discussion

By means of the statistical analysis of the questionnaire survey, the current status and problems of graduate students in life science disciplines about their data literacy are summarized. In addition, the results help us have a better understanding of their needs for data literacy education and provide evidence for libraries to improve data literacy education services. It is shown in Table 2:

Table 2. Suggestions based on the survey results

	Problems description	Suggestions
Basic knowledge about research data	Lack knowledge of discipline data management and sharing policies Lower identification of new scientific data software	Strengthen basic knowledge of research data management in discipline Embedded in the discipline, training in the field of research data management expertise Strengthen the study of data management and sharing policies, such as research funding agencies, Periodical Publishers, etc
Data management plan	Lack ability of understanding data management plan and utilizing of data management planning tools	Strengthen the understanding of research data lifecycle and data management plan. Development the ability of using data management planning tools
Data collection and documentation	Lack ability of discovering and reusing research data in data repository	Enhancement the ability to retrieve data sources such as data platforms, data repositories
Data processing and analysis	Unable to select and use appropriate data processing tools and software; Lack ability of assessing data quality	Strengthen the learning of the appropriate data processing and analysis tools. Introduction of the methods to control the quality of scientific research data
Data management and preservation	Lack ability of specifications metadata, data backup, data security measures	Improvement awareness of protecting data security, learning data backup strategy and measures
Data publishing and sharing research data ethics	Lack awareness of data publication and sharing	Enhancement data sharing awareness and choosing the appropriate way to publish data
Data rights and ethics	Lack awareness of data citing and the rights related to scientific research data	Learning knowledge of data citation and the rights of scientific data management

5 Data Literacy Education Design

Based on survey, the data literacy education model of teachers, students and curriculum is constructe (Fig. 1), the education content is based on the research data lifecycle and includes three levels of learning modes, that is, basic learning, advanced learning and promotion learning. At the same time, we have also designed data literacy education implementation scenes for graduate students in UCAS, and provide evidence for libraries to implement data literacy education services better.

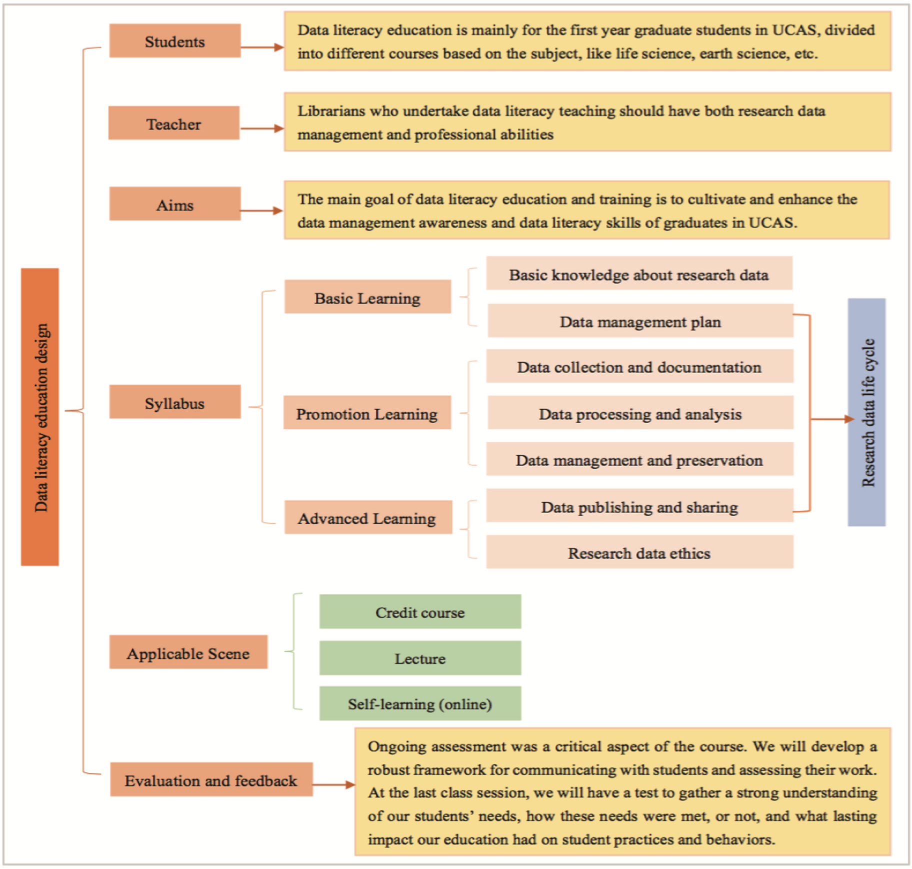
Data literacy education design
Students
Data literacy education is mainly for the first year graduate students in UCAS, divided into different courses based on the subject, like life science, earth science, etc.
Teacher
Librarians who undertake data literacy teaching should have both research data management and professional abilities
Aims
The main goal of data literacy education and training is to cultivate and enhance the data management awareness and data literacy skills of graduates in UCAS.
Syllabus
Basic Learning
Basic knowledge about research data
Data management plan
Promotion Learning
Data collection and documentation
Data processing and analysis
Data management and preservation
Advanced Learning
Data publishing and sharing
Research data ethics
Research data life cycle
Applicable Scene
Credit course
Lecture
Self-learning (online)
Evaluation and feedback
Ongoing assessment was a critical aspect of the course. We will develop a robust framework for communicating with students and assessing their work. At the last class session, we will have a test to gather a strong understanding of our students' needs, how these needs were met, or not, and what lasting impact our education had on student practices and behaviors.

Fig. 1. Data literacy education design

6 Conclusion

In this research, through conducting data literacy investigation among graduate students by means of a questionnaire survey, the current situation of their data literacy and the need to develop them are revealed. Based on the survey results, we developed a tailor-made course to help students improve their data literacy competencies. The results not only provide an implementation program for the improvement of graduate students in UCAS, but also provide a lot of useful information for assisting future librarians in incorporating data literacy skills into their services, especially for academic and research librarians to prepare and develop a data literacy course for their fields' graduate students.

References

1. Association of College and Research Libraries: Working Group on Intersections of Scholarly Communication and Information Literacy. Intersections of Scholarly Communication and Information Literacy: Creating Strategic Collaborations for a Changing Academic Environment. Association of College and Research Libraries (2013)
2. Carlson, J., Fosmire, M., Miller, C.C., Nelson, M.S.: Determining data information literacy needs: a study of students and research faculty. Portal: Libr. Acad. **11**(2), 629–657 (2011)
3. Carlson, J., Johnston, L., Westra, B., Nichols, M.: Developing an approach for data management education: a report from the data information literacy project. Int. J. Digit. Curation **8**(1), 204–217 (2013)
4. NECDMC. http://library.umassmed.edu/necdmc/index
5. Data Information Literacy. http://www.datainfolit.org/
6. DataTrain Project. http://www.lib.cam.ac.uk/preservation/datatrain/
7. MANTRA Research Data Management Training. http://mantra.edina.ac.uk/