INTEGRATED RESEARCH DATA MANAGEMENT CURRICULUM FOR PHD STUDENTS FROM THE PERSPECTIVE OF A LIBRARIAN

Svitlana Chukanova, PhD in Pedagogy,

National University of Kyiv-Mohyla Academy, Ukraine

***Abstract**: Open Science principles are difficult to imagine without proper research data curation and management as research data became the most essential source of modern research findings. That is why it is important to teach researchers and PhD students from the beginning of their academic careers. Research data management topic deserves special attention as it is a huge part of information management and information literacy practice. In terms of the PhD students' training curriculum, it is necessary to distinguish between RDM training for librarians and data curators and RDM training for researchers. The last one must be a simplified version of library specialists' raining and contain less library terminology as this area is new and complex.*

Keywords: *research data management, training, PhD studies, data curation*

INTRODUCTION

The course on Research Data Management (RDM) is an essential part of PhD studies in the modern educational context. The practice of research data management might be quite new to PhD students who are at the beginning of their academic careers and who just have started their journey in acquiring and developing transferable skills. P. Denicolo and J. Reeves indicate that among the list of transferable skills necessary for academic career for doctoral students it is worth mentioning "research information and data management" (Denicolo & Reeves, 2014, p. 56). These researchers stress the point that any research has as a basis particular data or information and thus doctoral students must know how to interact with it most effectively and efficiently (Denicolo & Reeves, 2014, pp. 56-57). The concept of information literacy is especially important in helping doctoral students to develop their research information management skills.

RESEARCH DATA MANAGEMENT: CLOSER LOOK

Ottawa University Library defines RDM as "…the active organization and maintenance of data throughout the research process, and suitable archiving of the data at the project's completion. It is an on-going activity throughout the data lifecycle" (What is research data management?). This statement means that RDM is a process which lasts during the research project and has several stages that corelate with research data lifecycle stages. It is important to mention that RDM is a cooperation between the national policy of research data processing, institutional

policy, library, and the researchers themselves as they are main figures in delivering the research project and gathering the data. RDM training can be performed with the use of different approaches and with the involvement of open resources and courses. The most known platforms for RDM training are Research Data MANTRA - free online course developed by the University of Edinburgh (Research Data MANTRA); DataONE Education Modules (Education Modules | DataONE, 2012); FOSTER (Foster). For example, let us regard closely one of the course structures - Research Data MANTRA incorporated in the table below.

Research Data MANTRA platform description. Openness: to everyone without registering and consists of elements as described in the following text. Module on research data in context with submodules: types of research data; why managing data is important; challenges of data in society. Module on data management planning with the submodules: good practice and responsible research; checklists and planning tools; funder compliance. Module on organising data which constitutes from: naming and re-naming conventions; file and code versioning; use of cloud collaboration tools. Module on file formats and transformation with such submodules: open and proprietary formats; compression; normalisation. Module on documentation, metadata, citation with the submodules: data documentation; using other peoples' data; forms and purposes of metadata; data citation as part of the scholarly record. Module on storage and security with such elements: backup and storage methods; password safety and encryption; strategies for long-term data security. Module dedicated to protecting sensitive data with the following elements: data protection legislation; ethical considerations and informed consent; safeguarding sensitive data. Module on sharing, preservation and licensing contains such submodules: formal and informal data sharing; preservation and trusted repositories; licensing and "open data". Data handling tutorials: Practice manipulating data in software analysis packages (SPSS, R, ArcGIS, NVivo) using open datasets with exercises in PDF (Research Data MANTRA).

These mentioned above course may be of use for library and information scientists and professionals, especially data librarians and data curators but they also will be a great guide for the researchers who want to learn more about RDM. Among the resources on RDM developed by different academic libraries and information centres it is worth to mention the following examples: Data management guidelines from Finnish Social Science Data Archive (Data management guidelines - Finnish social science data archive (FSD)); Additional resources accompanying a book "Managing and Sharing Research Data: a Guide to Good Practice" by Louise Corti, Veerle Van den Eynden, Libby Bishop, Matthew Woollard posted on SAGE publishing website (Managing and sharing research data: A guide to good practice | online resources, 2020). These resources explain various aspects of Open Science in general and especially Research Data Management. This list can be continued as modern libraries and data centres are

working on their guidelines, develop new MOOCs, and preparing training materials. By regarding all the resources and platforms mentioned earlier I suggest developing the following curriculum in RDM for Ph.D. students which would contain the topics on the notion of research data, what is RDM, working with data (including data organizing), data sharing, data preservation, and data ethics of Academic Integrity aspects in data management. Let us take a closer look at the example of one module which may be taught in terms of such a course.

THE EXAMPLE OF THE MODULE "THE NOTION OF RESEARCH DATA AND RESEARCH DATA LIFECYCLE"

This topic is dedicated to the basic notions of RDM and aimed to show the researcher what is research data and how it can be managed. The subtopics may be the following: what are research data; types and kinds of research data; different approaches to defining research data lifecycle; The DataONE data life cycle and it's components: plan, collect, assure, describe, preserve, discover, integrate, analyse. Students and researchers should bear in mind that there are diverse types and approaches to describe research data lifecycle, so, it is important for the instructor or lecturer to explain this variety and to choose one or a few lifecycle models for using in the course. Course participants must learn the connection between the lifecycle components to the activities provided during the data management and sharing process.

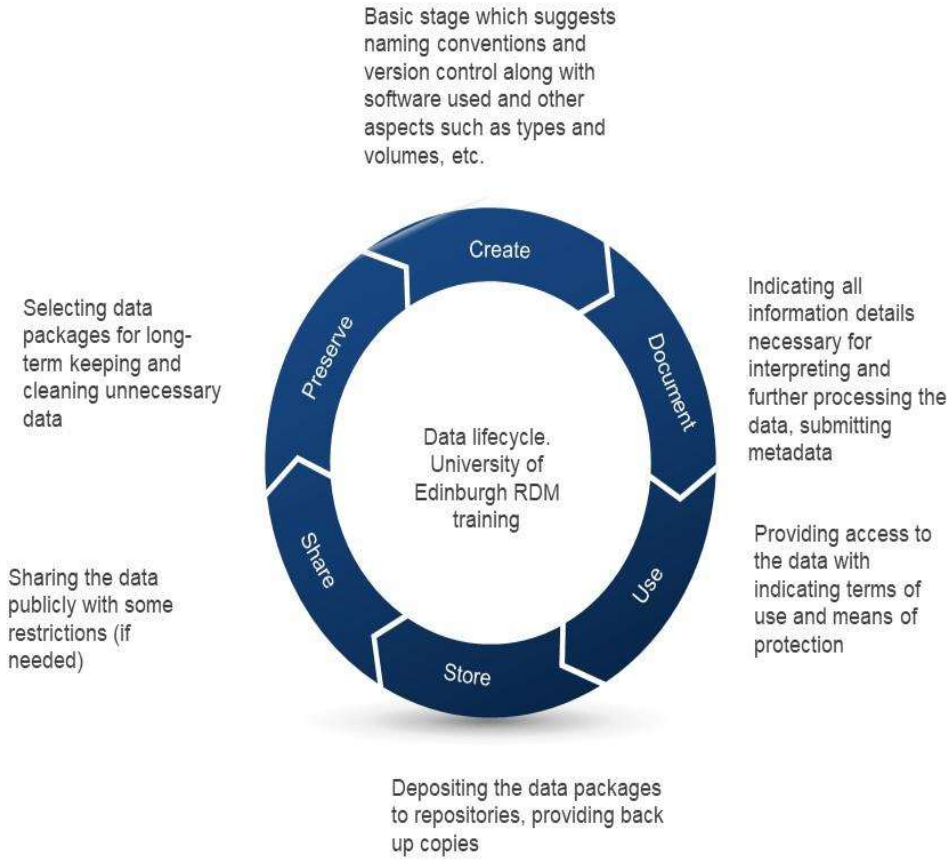
Research data notion is regarded by library scientists, data scientists as a basis for the research, interpreted facts which describe some idea or concept (Borgman, 2015, pp. 18-30). According to Vera Lipton: "Good data enables good science, and digital technologies provide the means for acquiring, transmitting, storing, analysing and reusing massive volumes of data…This is open innovation in science, or open science" (Research Data Management at CERN, 2020, p. 16). Thus, by talking about research data sharing we cannot but mention Open Science. In modern day world when Big Data grows popularity and scholarly communication upgrades faster and faster, we become deeply involved in the discussion of Open Data. It is possible to open data in different modes.

Suggested activities: 1**.** Compare two research data lifecycles (for example DataOne and Edinburgh University) and describe what did you find common and what are the differences between the chosen lifecycle stages? 2. Try to imagine the research data lifecycle of your research, describe what RDM action will you do at each stage? How would you define research data lifecycle? Will you include additional stages, or will you join a number of stages into one? Explain your point of view. 3. Analyse what resources will you need in performing RDM of your data?

EXAMPLES OF CONTENT FOR WORKSHOPS

This data lifecycle is developed by DataONE (Data Observation Network for Earth): **Plan**. At this stage, the researcher should provide the data description,

explain how the data will be managed and accessed. Planning is a particularly important stage for any research; **Collect**. This stage means that observations are made manually or utilizing different equipment and data are digitized and stored; **Assure**. This means that data are checked and verified thus has a quality; **Describe**. Data are described in detail using metadata to make data package usable and reusable; **Preserve**. This is vital stage of data processing and suggests depositing data into a data repository for long-term keep; **Discover**. Data is accessible for reuse in a future by other researchers or the author of the data to verify and validate the research or to build new research based on the previous; **Integrate**. The process of mixing data from several sources to construct one data package ready for analysis and processing; **Analyse**. Data can be analysed through different approaches, equipment, and methods (Education Modules | DataONE, 2012). For a better understanding of how research data lifecycle functions, it is useful to show another approach for comparison. With this aim, I suggest the research data lifecycle scheme from the University of Edinburgh RDM training.



The image is developed by S. Chukanova and used with the permission of R. Rice (R. Rice Rice & Southall, 2016, p. 78).

This data lifecycle is composed of the following components: "create, document, use, store, share, preserve" (R. Rice Rice & Southall, 2016, p. 78). The stage of creating implies the choice of types and formats, the use of software for

accessing data sets. These actions are remarkably similar to those indicated in the stage "planning" in the DataOne lifecycle. The stage of documenting resembles to stages "collect" and "describe" as we include the metadata aspect here. The stage of usage is quite alike stage "assure". Storing and preserving stages correspond with the stage "preserve", while the stage sharing is very similar to "discover". The stage of sharing is related to the stages "integrate" and "analyse" (Education Modules | DataONE, 2012).

Table 1: For the convenience of comparison, presentation in tabular format

DataOne Data Lifecycle	University of Edinburgh Data Lifecycle	Data aspects
Plan	Create	Formats, types, software, naming conventions, version control
Collect	Document	Metadata, describing the data
Describe		
Assure	Use	The restrictions of usage, verification, and validation
Preserve	Store	The depositing and long-term keeping
Integrate	Share	Reuse of the data or use of other authors' datasets
Analyse		
Preserve	Preserve	Keeping and depositing data (remarkably like store), taking a decision on what data to keep and what to retract.

By the end of the topic students will have their opinion on research data types and research data lifecycle relating to their studies. They will understand the importance of proper RDM actions taken on each step of the lifecycle.

SUMMARY

Research Data Management training for PhD students should contain modules which develop the following competencies: understanding of research data types and ability to distinguish between them, ability to categorise and classify research data, awareness that there are a lot of research data lifecycle schemes, ability to differentiate between research data lifecycle steps and what RDM actions are performed in terms of each step, ability to compose RDM plan with help of special open source tools, understanding of the requirements to RDM plans from different funding agencies, ability to prepare metadata according to a particular dataset, understanding the importance of open formats for better data sharing in the research community, understanding the principle of automatization in research data processing, understanding of file processing, ability to find data repository correspondent to the needs in data sharing, understanding the importance of personal data security and research integrity, understanding the importance of data packages citation, ability to use open tools for Data Science.

REFERENCES

1. Borgman, C.L. (2015). *Big data, little data, no data*. The MIT Press. https://doi.org/10.7551/mitpress/9963.001.0001

2. *Data management guidelines - finnish social science data archive (FSD)*. (б.д.). Finnish Social Science Data Archive (FSD). Available at: https://www.fsd.tuni.fi/aineistonhallinta/en/

3. Denicolo, P., & Reeves, J. (2014). *Developing transferable skills: Enchancing your research and employment potential*. SAGE.

4. *Education Modules, DataONE*. (2012). DataONE. https://old.dataone.org/education-modules

5. *Foster*. (б.д.). FOSTER. Available at: https://www.fosteropenscience.eu/

6. *Managing and sharing research data: A guide to good practice | online resources*. (2020). Online Resources. Available at: https://study.sagepub.com/corti2e

7. *Research Data Management at CERN*. In (Ed.), Open Scientific Data - Why Choosing and Reusing the RIGHT DATA Matters. IntechOpen. Available at: https://doi.org/10.5772/intechopen.91715

8. *Research Data MANTRA*. (б.д.). Research Data MANTRA. Available at: https://mantra.edina.ac.uk/

9. Rice, R., & Southall, J. (2016). *The data librarian's handbook*. Facet Publishing.

10. *What is research data management?* Library. Available at: https://biblio.uottawa.ca/en/services/faculty/research-data-management/what-research-data-management