Jun 14th, 9:35 AM - 10:50 AM

Doctoral Students' Research Data Management Competencies Based on the Quality of Their Data Management Plans

Jukka Rantasaari
Turku University Library Finland, jukka.rantasaari@utu.fi

Doctoral Students' Research Data Management Competencies Based on the Quality of Their Data Management Plans

Jukka Rantasaari
University of Turku
jukka.rantasaari@utu.fi,
https://orcid.org/0000-0001-5927-3781

Abstract

Many international, national, and institutional principles and policies as well as a growing number of funders and publishers recommend or mandate researchers to write data management plans (DMPs) or share the underlying research data upon which the research articles are based. At the core of these alignments is to enhance research transparency, reproducibility and reliability, and the reuse of research data by bringing the data findable, accessible, interoperable, and reusable (FAIR). To help researchers fulfill this task, they need education in research data management (RDM).

The goal of this preliminary paper is to find out the quality of the DMPs developed by doctoral students (DSs) in a 10-week, 3 ECTS credits multi-stakeholder Basics of Research Data Management (BRDM) course. Moreover, we aim to identify differences between DMPs in relation to background variables such as year, discipline, course track or other variables. The course is held in two multi-faculty research-intensive universities in Finland since 2019. In this ongoing study, 130 DSs' DMPs have been assessed and rated from 2020 and 2021 so far, using the criteria of the Finnish DMP Evaluation Guide (FDEG).

The quality of the DMPs appeared to be satisfactory. The differences between DMPs developed in separate years, course tracks or disciplines were statistically insignificant. However, DMPs that contained a data type specific classification (a data table) differed statistically highly significantly from DMPs without a data table. DMPs with a data table acknowledged better than DMPs without a data table the data handling needs of different data types and improved the overall quality of a DMP.

DMPs illustrated how well DSs had learned RDM competencies and how the course had furthered comprehension of the importance of sound data management practices to the integrity and reliability of the research, to the reusability of data, and to the reproducibility of the research.

> **This paper is a preliminary, abridged and limited version of the full article, which will be published later.**

1. Introduction

In 2016 Nature journal conducted a survey for researchers asking them "is there a reproducibility crisis in science?" (Baker, 2016). 52 per cent of 1 576 researchers

answered that yes, and the crisis is significant, meaning that the results of many scientific studies cannot be reproduced and are thus likely to be wrong (Baker, 2016).

Research data is all the information systematically acquired, processed, and analysed into new knowledge in academic research (Pryor, 2012). Consequently, data is a means to validate the research results: "a product of research and an essential part of the evidence necessary to evaluate research results, and to reconstruct the events and processes leading to those results" (Research Information Network, 2008). By organizing, storing, and preserving the data meticulously and sharing it linked with the research paper, other researchers and users can understand, verify, and reuse the data, and better reproduce the research. Thus, sound research data management (RDM) practices advance the integrity of data, reliability of research results, and reproducibility of research (e.g., Chiarelli et al., 2021). Of course, better data management alone may not be enough to secure the reproducibility. But if a researcher, from the start of the project, is committed themself to the transparent research process, and – when possible - open sharing of the data, they probably pay more attention to the sound and systematic collecting, processing, checking, and documentation practices throughout the research project enabling the sharing and reusing the data and reproducing the research.

For these reasons, during the second decade of the 2000s, many international, national, and institutional principles and policies and an increasing number of funders and publishers started recommending or mandating researchers to write data management plans (DMPs) and share data (e.g., Academy of Finland, 2019; "Amsterdam call for action on open science", 2016; European Commission, 2018a; 2018b; European University Association, 2017; National Science Foundation, 2011; UNIFI, 2016; Wellcome, 2017). However, to meet these recommendations and mandates, researchers need education, guidance, and support in RDM.

The aim of this preliminary paper is to find out doctoral students' (DSs) RDM competencies as they are manifested in the DMPs developed by DSs during the Basics of Research Data Management (BRDM) course held in the University of Turku (UTU) since 2019, and in UTU and Åbo Akademi University (ÅAU) since 2020. From the year 2020 participants have prepared a research plan and a DMP for their own PhD project in the course. For that purpose, 130 DMPs from the years 2020 and 2021 have been assessed and rated so far. We aim at answering the questions:

RQ1: What is the quality of the doctoral students' DMPs as rated according to the criteria of Finnish DMP Evaluation Guidance (FDEG)?
RQ-2 What kind of differences were found between DMPs regarding the year, discipline, course track, or other background variables?

2. Methods

2.1 BRDM Course

BRDM course was developed and implemented in 2019 at the University of Turku (UTU), the third-largest research-intensive university in southwestern Finland with eight faculties, five independent units, and 21 000 students including 2000 doctoral students and 3300 staff members. Before the course, the author interviewed 35 doctoral students, supervisors, and biostatisticians in UTU to learn the perceived importance of RDM competencies and doctoral

students' current competencies (Rantasaari, 2021; Rantasaari & Kokkinen, 2019). Data management planning, documentation of data processing, and managing IPR and contract issues contained the most profound skills gaps. However, participants also lacked knowledge of different issues throughout the data lifecycle. Therefore, the author, with the leader of UTU's biostatistician team, set up a working group and invited researcher-teachers from different faculties, a grant writer, data librarians, lawyers, a data security officer, and an IT computing specialist to plan and teach a course on RDM for doctoral students (DSs) and postdoc researchers. In 2020 we extended the course to Turku's other university - Åbo Akademi University (ÅAU) - the only Swedish language multi-faculty university in Finland with 5500 students including 700 doctoral students and 1100 staff members. In this preliminary paper, we focus on the quality of DMPs developed by DSs.

In table 1 are illustrated the 10-week, 3 ECTS credits multi-stakeholder Basics of Research Data Management (BRDM) course structure, contents, and teachers. The course consists of four tracks, introductory lecture, seven modules, voluntary Q&A session, and a final assignment. The teachers are academic and research support professionals. The idea behind the four-track-based division is that the type of data management actions needed and applied depend partly on the type of the data, partly on research methods, and partly on discipline.

In module one, each participant developed their own research plan. After that, based on their research plan, they started to write a DMP in the DMPTuuli tool[1]. Each modules' pre-class assignment was to write a draft of the relevant section of the DMP while the post-class assignment was to update the section, informed by the modules' workshop. At the end of the course, everyone returned their DMPs, and gave a structured, anonymous peer-review report of another participant's DMP. Finally, we gave a general level and personal feedback to everyone. Participants were recommended to use the Finnish DMP Evaluation Guidance (FDEG) (Aalto et al. 2021) as an aid to prepare their DMP and assess another participant's DMP. FDEG was developed and implemented by the Finnish Tuuli working group chaired by the author during the Spring of 2021.

Table 1: The structure, contents, and teachers of the four-track BRDM course

---

[1] https://dmptuuli.fi/

CLINICAL HEALTH SCIENCES	SURVEY RESEARCH	QUALITATIVE RESEARCH	NATURAL SCIENCES	TEACHERS
Introductory Lecture: Background and concepts; Characteristics of a high quality research plan; Course practicalities				Head of library services; Data librarian; Grant writer
Research plan: Objective; design; implementation; expected results	Research plan: Objective; design; implementation; expected results	Research plan: Objective; design; implementation; expected results	Research plan: Objective; design; implementation; expected results	Researchers; Lectors; University teachers
Data management plan (DMP): data life cycle; DMP-tool; roles; resources; metadata;documentation				Self-study module; Data librarian
IPR, agreements and licenses (in Finnish)		IPR, agreements and licenses (in English)		Head of legal affairs unit; Assistant legal counsel; Data librarian
Data privacy: privacy notice; risk analysis; (in Finnish)		Data privacy: privacy notice; risk analysis; anonymization (in English)		Data protection officer; Data archive specialist
RedCap: building a form based database	RedCap: building a survey form	NVivo: organizing and coding data	RedCap: building a form based database	Head of biostatistician team; Lector
Data storage, protection, processing, describing and IT Service solutions				IT system architect
Data preservation, sharing and citing; Open data repositories				Data librarian
A voluntary Q&A Session				Module teachers
DMP: returning and peer-reviewing	DMP: returning and peer-reviewing	DMP: returning and peer-reviewing	DMP: returning and peer-reviewing	Head of library services; Data librarian
a general level feedback on DMPs				Head of library services

2.2 The assessment, rating, and analysis of the DMPs

By participating BRDM, doctoral students accepted that DMPs and other anonymised course assignments can be used for research and development of the curriculum. The assessment and rating of the DMPs developed in the course are based on the idea that DMPs illustrate how well DSs have learned the concepts and principles of the different aspects of RDM taught in BRDM, and to what extent they can apply the principles in their own data management planning. Thus, DSs' DMPs were assessed and rated according to FDEG's three-point criteria (Aalto et al. 2021). DMPs and each of their sections were rated as excellent (2 points), satisfactory (1 point) or poor (0 points). All the DMPs were read and assessed by the author of this study. After assessing a DMP, the author read it second time using another participant's peer-review report as a reference. If the score of the first and second read differed markedly, the author read and assessed DMP third time.

For analyzing the results of the ratings, we used SAS JMP Pro 16 statistical software to produce descriptive and inferential statistics with medians, custom quantiles, and Wilcoxon signed rank tests. A significance level of 0.05 (two-tailed) was used.

3. Results

3.1 Participants

In total, 189 DSs and 28 postdoc researchers participated the BRDM course in 2020 and 2021. In this preliminary paper we focus on the DMPs created by DSs. Of all the DSs, 69 per cent (130) completed the full course and returned their DMP, whereas 31 per cent (59) completed on average four modules according to their interest in the subjects. Fully completed DSs came from social sciences, business, and economics (n=52); health sciences (36); science and engineering (31); and humanities, psychology, and theology (11).

3.2 Data management plans in 2020 and 2021

In terms of FDEG's three-point criteria, the difference between DMPs developed in 2020 and 2021was statistically insignificant. The median of all DMPs prepared by DSs in 2020-2021 courses (n=130) was 1 (Q1:0.97, Q3:1.04). Likewise, the differences between DSs' DMPs in different course tracks or disciplines were statistically insignificant. The green bars in figure 1 illustrate the rating of the sections of all DMPs. The median values of most of the sections are situated very close to 1 except Section 4.1 "Storage and security" (1.63) and Section 6.2 "Budgeting and resourcing" (0.69).

However, DMPs that contained a data type specific classification (a data table) differed statistically highly significantly (p<0.0001, Wilcoxon Rank Sum Test) from DMPs without a data table. A data table which was recommended in FDEG (Aalto et al. 2021) to be included in the Section 1.1 of a DMP, contained a clear table or a list of all reused, collected, and produced data types with their formats and volumes handled in the project (Table 2).

Table 2: An example of a data table

Data type and source	File format	Personal or sensitive data	Ownership and agreements	Metadata documentation	Storage during project	Opening or publishing data after the project	Long-term archiving	Estimated size
Lab notes (Data produced)	.doc .txt .pdf	Yes. Subject to IPR check	PI and group	Programme generates metadata by itself	Electronic lab notebook (eLabJournal)	Project team	Discarded after 15 years	< 10 MB
RNA sequences (Data produced)	raw: FASTA, BAM, .xlsx	no	PI	Readme.txt	Shared network drive hosted by UTU and backed up every day (Seafile)	European Nucleotide Archive	no	< 1 GB
MRI images (Data reused)	DICOM, .nii, .tiff	Yes, record keeper: xx	PI	Readme.txt	Database x at TYKS, backup	NITRC after anonymization	no	< 1 GB
Question-naire forms (Data collected)	Paper forms	Yes, record keeper: xx	PI	Readme.txt	Locked filing cabinets in PI office	No, metadata open in Zenodo/Etsin	Discarded 5 years after publication	
…								

DMPs with a data table contained more overarching and detailed information of the data management characteristics and actions concerning e.g., origin, format, personal data, sensitivity, license, metadata, collaborators' different rights of usage, and volume than DMPs that did not contain a data table (Figure 3). The median of the DMPs with a data table (53 %, n=67) was 1.08 (Q1:1.02, Q3:1.42) and without a data table (47 %, n=63) 0.97 (Q1:0.78, Q3:0.99) (Figure 1).

DMPs with a data table received the highest median rating in all sections. All the sections except Section 4.2 "Related data security policies" with a statistically insignificant difference (p=0.06) and Section 6.2 "Budgeting and resourcing" with a statistically significant difference (p=0.047), differed statistically highly significantly between DMPs with and

without a data table: Section 1.1 "Data description" (p<0.0001), Section 1.2 "Data quality" (p<0.0001), Section 2.1 "Legal issues" (p=0.0002), Section 2.2 "Rights management" (p=0.0004), Section 3.1 "Documentation and metadata" (p=0.003), Section 4.1 "Storage and security" (p=0.003), Section 5.1 "Data sharing" (p=0.001), Section 5.2 "Preservation" (p=0.0007), and Section 6.1 "Roles and responsibilities" (p=0.007).
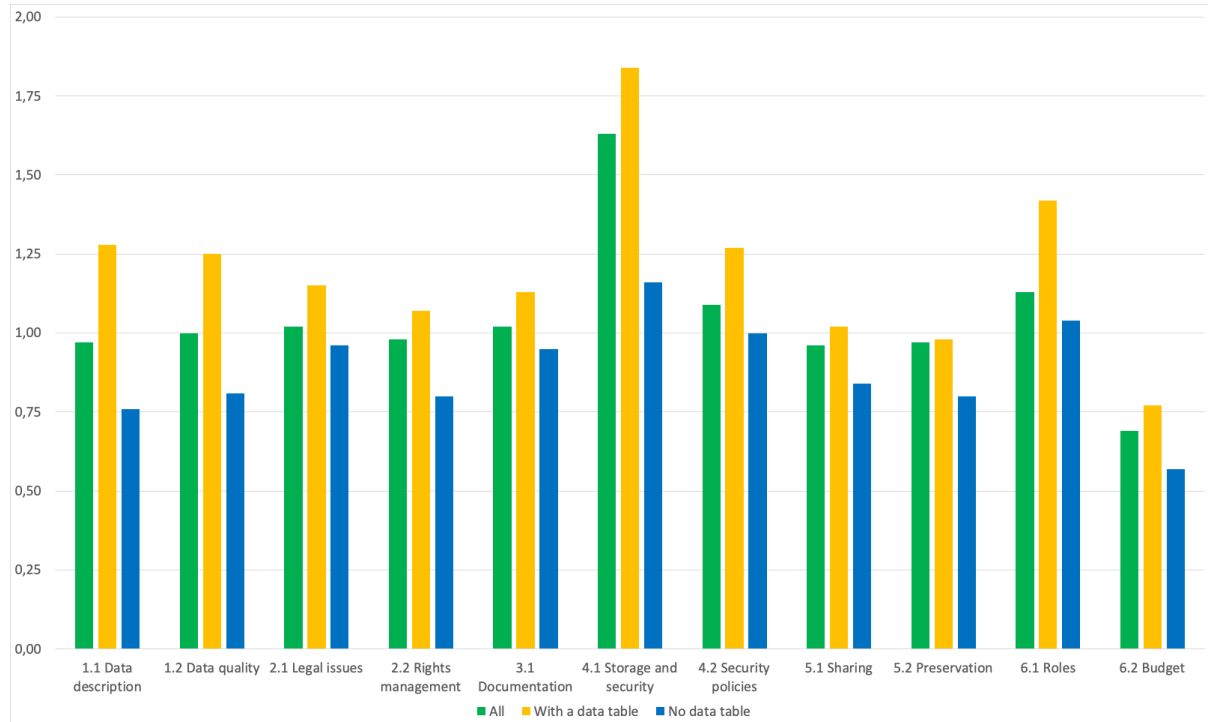

Figure 1: Medians of the sections of all DSs' DMPs (n=131) (green) and DMPs with (yellow) and without (blue) a data table in BRDM 2020 and 2021 courses. The titles of the sections in full: 1.1 Data description; 1.2 Data quality; 2.1 Legal issues; 2.2 Rights management; 3.1 Documentation and metadata; 4.1 Storage and security; 4.2 Related data security policies; 5.1 Data sharing; 5.2 Preservation; 6.1 Roles and responsibilities; 6.2 Budgeting and resourcing

4. Limitations

Because of the limited number of DMPs analyzed so far, we cannot generalize all the results of our study and the factors affecting them outside of the research group. Furthermore, we cannot know to what extent DSs will comply the actions they have described in their DMPs such as documentation, quality control, and sharing and preserving the research data or metadata. However, 130 returned DMPs will reveal valuable, indicative information of DSs' competencies, the impact of the education on competencies, and further learning needs in RDM.

5. Discussion and Lessons learned

The median rating of the DMPs prepared in 2020-2021 courses was satisfactory. The median values of the DMPs' sections are situated very close to 1 with three exceptions: Section 4.1 "Storage and security" is rated nearly at an excellent level because almost all the storage solutions offered by the University of Turku and Åbo Akademi University fill the need of

safe and secure storage for any data types. Another two exceptions from the medium value of 1 can be seen in the sections 6.1 "Roles and responsibilities" and 6.2 "Budgeting and resourcing". The plans how to share the RDM roles and responsibilities between collaborators in a research project are between satisfactory and excellent whereas the plans how to resource and budget the practices are between poor and satisfactory. On one hand, it may not sound good that DMPs' authors could not evaluate the timely and financial resources needed for data management actions. However, many authors emphasized that they saw research data management as an intrinsic part, not an extra burden of research process.

DMPs that contained a data table received significantly better ratings compared to DMPs that did not contain a data table. The reason for the difference was that DMPs with a data table acknowledged better than DMPs without a data table the data handling needs of different data and improved the overall quality of a DMP with more granular description of data types, documentation, metadata, and actions needed to handle different kinds of data.

Based on the analyses so far, the quality of the DMPs did not differ according to DSs' discipline or course track. There are three potential reasons for undiscovered differences. First, teaching and instructions in BRDM course specifically highlighted sound and FAIR principles and practices in RDM instead of existing cultures and practices followed in different disciplines. Second, we have not yet conducted advanced content analysis of the DMPs' key characteristics that can reveal differences between disciplines, used research methods and data types. Third, the impact of discipline, research method and data type on differences in RDM practices can often be interweaved.

Next, the study at hand will be extended first by a literature review of previous studies on DMPs' assessment, second by assessing and rating the DMPs from BRDM 2022 course, and third by assessing and rating postdoc researchers' DMPs from BRDM 2020, 2021, and 2022 courses. Moreover, content analysis of all DMPs between BRDM 2020 and 2022 courses will be elaborated: We will examine how the key themes of RDM such as sound documentation and metadata, safe and secure storing, sharing of data or metadata, preservation of data, and legal and ethical rights management have been applied in the DMPs. Moreover, we will find out if there are differences between disciplines, course tracks or other background variables in relation to these key themes.

References

Aalto, S., Ahokas, M., Friman, J., Fuchs, S., Korhonen, T., Kuusniemi, M. E., Laakso, K., Lennes, M., Manninen, S., Ojanen, M., Rantasaari, J., Virtanen, M. E., Xu, Q. (2021) *Finnish DMP evaluation guidance*. [Working paper]. Zenodo. https://doi.org/10.5281/zenodo.4729831

Academy of Finland. (2019). *The Academy of Finland's funding terms and conditions 2019-2020* (Vol. 1, Issue September 2019). https://www.aka.fi/globalassets/10rahoitus/liiteet/rahoitusehdot_2019-2020_en.pdf

Amsterdam call for action on open science. (2016, April 5). In W*ikipedia*. https://en.wikipedia.org/wiki/Amsterdam_Call_for_Action_on_Open_Science

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454. https://www.nature.com/articles/533452a

Chiarelli, A., Loffreda, L., & Johnson, R. (2021). *The art of publishing reproducible research outputs: Supporting emerging practices through cultural and technological innovation*. [Report]. Zenodo. https://doi.org/10.5281/ZENODO.5521077

European Commission. (2018a). *OSPP-REC: Open science policy platform recommendations*. OSPP-REC. https://doi.org/10.2777/958647

European Commission. (2018b). Commission recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information. *Official journal of the European Union* L 134, p. 12–18 . http://data.europa.eu/eli/reco/2018/790/oj

European University Association. (2017). *Towards full open access in 2020: Aims and recommendations for university leaders and national rectors' conferences*. http://www.eua.be/Libraries/publications-homepage-list/towards-full-open-access-in-2020-aims-and-recommendations-for-university-leaders-and-national-rectors-conferences?platform=hootsuite

National Science Foundation. (2011). *Grant proposal guide* (NSF 11-1). https://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp

Pryor, G. (2012). Why manage research data? In G. Pryor (Ed.), *Managing research data* (p. 224). Cambridge: Facet Publishing. doi:10.29085/9781856048910

Rantasaari, J. (2021). Doctoral students' educational needs in research data management: Perceived importance and current competencies. *International Journal of Digital Curation*, *16*(1), 1-36. https://doi.org/10.2218/IJDC.V16I1.684

Rantasaari, J., & Kokkinen, H. (2019, June 23-27). *Closing the skills gap: The basics of the research data management (BRDM) course: Case University of Turku*. [Conference presentation]. The fortieth IATUL conference, Perth, Australia. https://docs.lib.purdue.edu/iatul/2019/fair/5/

Research Information Network. (2008). *Stewardship of digital research data: A framework of principles and guidelines*. Retrieved from http://shapingthefuture.pbworks.com/f/Stewardship+of+digital+research+data+by+RIN.pdf

UNIFI. (2016). *Open science and data: Action programme for the Finnish scholarly community*. https://www.unifi.fi/wp-content/uploads/2019/04/UNIFI_Open_Science_and_Data_Action_Programme.pdf

Wellcome. (2017, July 10). *Data, software and materials management and sharing policy*. https://wellcome.ac.uk/funding/guidance/data-software-materials-management-and-sharing-policy