TASK QUARTERLY vol. 25, No 4, 2021, pp. 499–509

RESEARCH DATA MANAGEMENT IN THE CONTEXT OF RESEARCHERS' NEEDS AND OPEN SCIENCE ADVANCEMENTS IN POLAND. OVERVIEW OF THE TRAINING PROGRAMME OF INTERDISCIPLINARY CENTRE FOR MATHEMATICAL AND COMPUTATIONAL MODELLING, UNIVERSITY OF WARSAW

NATALIA GRUENPETER

Interdisciplinary Centre for Mathematical and Computational Modelling
University of Warsaw
Adolfa Pawińskiego 5A,
02-106 Warszawa, Poland
e-mail: n.gruenpeter@icm.edu.pl

(received: 16.12.2021; revised: 16.12.2021;
accepted: 21.12.2021; published online: 30.12.2021)

Abstract: The article presents conclusions from research data management (RDM) trainings conducted since 2015 as part of the activities of the Open Science Platform, an initiative of the Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw. The aim of the article is to present a training programme and conclusions based on 5 years of experience (in the years 2015–2020 more than 40 workshops and trainings were carried out, face-to-face and online). The frequently asked questions and problems raised by the participants will be discussed. Also, issues addressed during the trainings will be placed in a broader context of changes in scientific communication in Poland, including, in particular, the implementation of openness policies and the development of new research data repositories (as part of the Disciplinary Open Research Data Repositories project). The article will also suggest possible improvements and developments in RDM training.

Keywords: research data management, scientific communication, data repositories, Open Science Platform

DOI: https://doi.org/10.34808/tq2021/25.4/s

1. Introduction

In recent years, the issue of research data management (RDM) has become a subject of increased interest from Polish scientific institutions, as well as the researchers, librarians, and administrative staff employed therein. Guidelines for completing the data management plan (DMP) issued by the National Science Centre in 2019 and the Open Access Policy adopted a year later were one of the most important factors affecting this state of affairs [1]. Although the latter document implements Plan S, which, as in its current shape, concerns only articles published in scientific journals [2], it also includes some provisions on research data. To some extent, it complements the guidelines for completing the DMP: "Where possible, the underlying data (basic datasets) related to published articles should be placed in an open repository, based on a Creative Commons Public Domain license (CC0 license), with the data citation standards laid down in the Declaration of Data Citation Principles by FORCE 11, as well as the principles specified in the TOP Guidelines. All published metadata must meet the guidelines laid down by OpenAIRE (https://zenodo.org/record/6918.XqfeHf0zbIU) and include a note about the project funding resources (National Science Centre, project number)" [3]. These provisions were introduced on 31 May 2020 and, as a first step, were applied to research projects, scholarships, fellowships, and research activities recommended for funding, starting from the call for proposals announced on 16 June 2020, for which funding agreements were supposed to be signed after 1 January 2021. Currently, the issue of open research data is also taken into account in "The Open Data and Re-use of Public Sector Information Act" that was promulgated in the Journal of Laws of the Republic of Poland in September 2021 [4]. It implements Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the reuse of public sector information, which to some extent also concerns research data. In addition, the Council of Ministers of the Republic of Poland has adopted this year the open data programme for 2021–2027 which covers key issues related to data sharing and management. Within the framework of Objective 4, which was defined as "stimulating the reuse of cultural resources and scientific data", an increase in the availability of scientific resources was announced, inter alia, by updating the national open access policy [5].

Appropriate RDM planning and providing open access to research data in accordance with the FAIR principles and the "as open as possible, as closed as necessary" principle were already required by the European Commission in projects implemented under the Horizon 2020 programme (with the possibility of opt-out in justified situations). This requirement is also in place for Horizon Europe, a new framework programme for research and innovation (RI) for 2021–2027. Documents published in May 2021 described the RDM requirements as follows: "Likewise, it should be ensured that beneficiaries provide open access to research data following the principle 'as open as possible, as closed as necessary' while ensuring the possibility of exceptions taking into account the legitimate

interests of the beneficiaries. More emphasis should, in particular, be given to the responsible management of research data, which should comply with the principles of 'findability', 'accessibility', 'interoperability' and 'reusability' (the 'FAIR principles'), in particular through the mainstreaming of data management plans" [6]. In practice, this means that responsible RDM in accordance with the FAIR principles should be planned and implemented in all projects that produce research data.

Also, Science Europe, the association representing major public organisations that fund or perform research in Europe, actively promotes the sharing and re-using of FAIR research as well as proper RDM planning. The organisation supports open sharing and reusing of research data as a means for generating impact and facilitating the verification of research findings and scientific progress. The involvement of Science Europe in this area includes, inter alia, promoting the management of research data in accordance with the FAIR principles, as well as the development of a harmonised framework and guidance for data management, preparation of data management plans and selection of trustworthy repositories. This resulted in guidelines [7] adopted and supported by many organisations that finance research, including the European Research Council and the European Commission [8], as well as the National Science Centre in Poland [9].

The aforementioned initiatives, as well as many other initiatives working towards open science, put a great emphasis on the objectives and benefits of proper RDM and data sharing, in accordance with the "as open as possible, as closed as necessary" principle. Appropriate RDM strategies and solutions not only improve research performance and enable verification of the results but also facilitate future re-use of research data. Benefits can be identified both from the point of view of the research performing and research funding organisations and from the point of view of researchers conducting research projects. In addition, open research data may be used by other researchers, entrepreneurs, journalists and experts in various fields, public and local administration staff members and non-governmental organisations, as well as by all citizens concerned. The basic assumptions and benefits of open research data are set out, inter alia, in the Sorbonne Declaration signed in January 2020 during the Research Data Rights Summit by representatives of networks of research-intensive universities from around the world. The document highlights the role of the integrity of ongoing research in building and strengthening the public confidence in science and the role of open data both for science itself and for social and economic development [10]. Furthermore, in January 2020, Science Europe published a document summarising the experience of the member organisations implementing research data policies and identifying good practices at three stages: policy development, communication and implementation. Raising awareness and understanding of proper RDM is the objective of communication activities that have the chance to fill the gap diagnosed by Science Europe: "Researchers are still often reluctant to dedicate time to the setting-up of a DMP rather than to their actual research project. Clearer

communication on the opportunities, usefulness and relevance of good RDM and the underlying concepts of Open Science, in general, and DMPs, in particular, is therefore needed. The development of DMPs should be seen as just another step in the process of conducting research" [11]. Communication understood as providing the necessary policy information also includes trainings. The authors of the document recognise researchers' needs: "Many scientists are expressing the need for more detailed information and training on RDM issues and for more support in the preparation of DMPs towards their home institutions or funding organisations. This is especially the case as far as understanding the legal aspects of RDM is concerned" [11]. Trainings are therefore an important part of open science activities.

2. Research data in the activities of the Open Science Platform

The Open Science Platform team operating at the Interdisciplinary Centre for Mathematical and Computational Modelling at the University of Warsaw has been organising and conducting trainings in research data management since 2015. These activities are part of the broadly defined tasks and projects such as developing the infrastructure for open science, supporting researchers, scientific institutions, and publishers in open sharing of research results, as well as monitoring and analysing developments in scientific communication. Activities in the indicated areas complement and reinforce each other, which allows comprehensive recognition of many phenomena related to open science.

A comprehensive approach to RDM and data sharing lies at the very foundation of the activity in question. On 28–29 May 2015, the international conference "Open Research Data: Implications for Science and Society" was organised in Warsaw by the Open Science Platform. It provided a forum for a broad debate on all issues related to open research data: policies, strategies, incentives, tools, and methodologies, as well as opportunities and challenges in the reuse of data [12]. In the autumn of the same year, the Repository for Open Data (RepOD) was launched. To this day, this general-purpose research data repository for open sharing of research data created, collected, or annotated for scientific research can be used free of charge by researchers from all scientific institutions and disciplines. The RDM trainings started in December 2015, following the launch of the repository. Their aim was to raise awareness of the need for RDM and to promote an active approach to research data at each stage of a research project. During the training sessions, the requirements of various institutions in terms of data sharing were presented, legal aspects of work with data were discussed, and the available technical solutions, i.e., repositories for data sharing, were demonstrated. Moreover, in 2016, the Towards Open Research Data in Poland report was published the aim of which was to initiate a broad discussion on open research data in Poland. The report consisted of two parts – an analysis of legal conditions for data sharing and a presentation of the results of a survey

conducted among Polish researchers. The purpose of the study was to "explore the phenomenon of academic data sharing in Poland (experiences with and attitudes towards the issue, enablers and obstacles, level of knowledge about legal and technical aspects)" [13]. The report was complemented by recommendations addressed to various stakeholders: research funding organisations, universities and research institutes, publishers of scientific journals, researchers, and relevant government departments. The need for training is noted in Section 6, which contains general recommendations for all groups.

3. Basis for conclusions

The conclusions presented in the following section of the article are based on the RDM training sessions conducted by members of the Open Science Platform team in 2015–2020. The five-year perspective provides the opportunity to observe changes in the broadly understood scientific communication as well as the problems, doubts, and comments formulated by the training participants. To some extent, they reflect the needs of researchers in terms of specific solutions, such as the available infrastructure and competence, such as knowledge of the FAIR principles or good practices for research data sharing. The article takes into account the discussions held during the training sessions and post-training surveys completed by the participants at the end of the workshops. The training sessions were conducted as part of the Open Science Platform and the National Open Access Desk OpenAIRE (NOAD) activities, and at the invitation of various scientific institutions and academic libraries from all over Poland. At the end of 2019, a series of training sessions began as part of the Disciplinary Open Research Data Repositories project, aimed at enhancing access to academic resources by making them available in open research data repositories, as well as improving the quality of shared data and metadata, and to facilitate their better use [14]. In 2020, the training courses were organised online due to the restrictions introduced throughout Poland related to the epidemic situation. The trainings were free of charge and open to all the concerned; only registration was required. They were attended by researchers, academic librarians, and administrative staff.

The training sessions were conducted in accordance with the plan developed by the Open Science Platform team, which was slightly modified over the 5 years, usually to update information on the openness policies, relevant projects, and initiatives or the available infrastructure and services. The training programme consisted of two parts. The first one concerned practical aspects of RDM, such as preparation of data for sharing in accordance with the requirements of research funding organisations, the FAIR principles, and good practices. The second, devoted to the legal aspects of research data management, concerned research data as the subject of legal regulation, including the intellectual property law, the General Data Protection Regulation and commercialisation of research results, as well as the use of free licenses (Creative Commons and Open Data Commons).

In the face-to-face training, both parts were supplemented by workshop activities. In the first part, this was an analysis of selected DMPs. Depending on the conditions and number of participants, the activity was based on individual work or group work, followed by discussion. In the legal part of the training, the activity included a role-playing method that took into account the perspectives of researchers and representatives of different types of research institutions, such as research funding organisations, research performing organisations and companies participating in research and development consortia.

The most important changes in the programme were introduced in the training sessions carried out in 2019 and 2020 as part of the Disciplinary Open Research Data Repositories project. The changes included adding information on the infrastructure created as part of the project, i.e., three data repositories: the new version of the Repository for Open Data (RepOD), Social Data Repository (RDS) and Macromolecular Xtallography Raw Data Repository (MX-RDR). On the other hand, some organisational changes resulted from the online form of training. The webinars carried out since March 2020 are shorter than the face-to-face workshops and are based on the lecture format. Question and answer sessions are held via text chat. Between 2015 and 2020, 42 training courses were held, most of which (more than 30) took place in the last two years. Taking into account the increased number of meetings and participants, as well as the timeliness of the problems and comments submitted by the participants, this period will be further analysed. Between 2015 and 2018, 15 training sessions took place, all of which were face-to-face workshops organised for circa 15–20 participants. In 2019, 15 training sessions were held , both in Warsaw and in other cities in Poland (Gdańsk, Kraków, Katowice, Lublin, Łódź, Poznań, Wrocław, Toruń), more than half of which were organised in cooperation with other institutions from all over Poland, most often with academic libraries. The increasing number of training sessions resulted from the increased interest in the subject of RDM which followed the National Science Centre decision to introduce a research data management plan as an appendix for grant proposals. In the autumn of 2019, training sessionsorganised as part of the Disciplinary Open Research Data Repositories project began and were continued in the next year in the online form. In 2019, more than 250 people participated in face-to-face the online formin total, and in 2020, more than 600 people participated in online the online form.

4. Frequently asked questions and comments

Taking into account the discussions with participants and the remarks made in the evaluation surveys in 2019, a clear division into the first and second half of the year can be indicated. Between April and June 2019, the participants most often raised two issues in the evaluation surveys (89 of them were filled in during the period in question). First, it was pointed out that the broad and multifaceted topics taken up during the trainings require additional, in-depth and specialised training modules, in particular on legal aspects (e.g.

intellectual property law, clearing copyrights) and practical aspects related to the collection of data (project documentation, anonymization of data, work with raw data and pre-developed data). Second, there has been interest in "real-life" examples and case studies, often without any further clarifications. Taking into account the scope of the topics undertaken, as well as the discussions during the trainings, it can be assumed that examples interesting for participants could include well-managed and open datasets (in particular the organisation of the data, various file formats, documentation and metadata) and analysis of the legal conditions for making specific datasets available. The aforementioned directions for developing the training offer also answer to some extent the need for discipline-specific training, which was also indicated in the surveys. This is an issue that has very often appeared in the feedback provided by training participants throughout the period in question. In addition, individual comments concerned data journals, perspectives of PhD students and young researchers , and more specific practical issues, such as instructions for depositing data in repositories. During the workshops that were organised between April and June 2019, the issue of data management plans did not appear on the first plan, and the participants did not raise any additional questions regarding the preparation of such documents. One of the exercises focussed on the analysis of data management plans shared in the DMPonline service [15]. English-language documents were discussed in terms of the consistency of the information about the research data, the adequacy of the solutions identified in the context of a short description of the study, the precision and clarity of the wording and, where possible, the degree of reference to good data sharing practices and standards. Concerning the exercise, the participants indicated their willingness to read and discuss DMPs written in Polish.

By the second half of 2019, this topic had already become the dominant theme of discussions conducted during the training sessions and comments made in the surveys. This was mainly due to the mandatory data management plan introduced by the National Science Centre as an appendix to the project funding application. Information on this topic was included in the communication on the NCN Council meeting (17–18 April 2019) [16]. The application period for calls with mandatory data management plans was from June to September 2019.

The need to expand on the topic of DMPs and to thoroughly discuss the guidelines of the National Science Centre was the most frequently reported comment in the surveys completed by training participants between October and December 2019 (participants completed 140 evaluation surveys during this period). Comments in this respect concerned, for instance, the example of an "ideal plan" or "correctly prepared DMP template". The funder's expectations regarding both the preparation of the document itself and its implementation, in particular, proper and secure data storage and archiving, as well as open data sharing. The participants' questions also concerned formal aspects, including the assessment of data management plans and their implementation, the possibility

of changing and updating the document during the project period, and the level of detail required in the plan, especially in the context of the limited number of characters on the application form. The most frequently asked questions were forwarded to the National Science Centre which responded and published the answers on its website [17].

In addition, during this period many questions and comments concerned institutional policies, services and other solutions that may affect research data management and sharing, such as the rules of intellectual property management, commercialisation, research data storage services, and institutional repositories. In the second half of 2019, many face-to-face training sessions were organised in cooperation with other institutions, but this topic was not included in the training programme. To some extent, it was raised in the legal part, however, not in a way that would take account of the full array of the regulations and solutions adopted in the specific institution. Furthermore, the participants submitted comments similar to those already discussed, in particular on data management and sharing practices adopted in various disciplines. In 2020, the training sessions were conducted remotely, which to some extent limited the possibility of carrying out conversations with participants. They were replaced by communication through a chat which allowed everyone to ask rather simple questions. Participants were also offered to fill in the same questionnaires as those used during the face-to-face training sessions. The questions raised during the online training sessions in 2020 most often referred to specific issues related to legal aspects of handling research data, the National Science Centre requirements, and the operations of data repositories. Also, there were questions and comments of a more general nature concerning the current shape and future of scientific communication, the importance of managing and sharing research data, and the researchers' attitudes towards open science practices, including their reluctance to make data available, the lack of an appropriate system of incentives and rewards, and concerns about the use of data in a way that would not be compatible with the license or would be unfair. The issues of scientific integrity and ethics, such as proper data citation, have very often been linked to legal issues, such as the practical implications of liberal licenses that enable the reuse and distribution of content. This was one of the most frequently discussed issues in the legal part of the training. Moreover, the legal issues most frequently asked about by the training participants included:

- legal aspects of the processing and managing of personal data, obtaining and documenting informed consent from participants, anonymization and pseudonymization under the EU data privacy rules;
- "ownership" of research results (both research data and other tools and instruments – software, algorithms, protocols, models, workflows, electronic notebooks and others), taking into account various models of collaboration, such as research consortia or industry and business partnership;
- the use of data stored and processed by commercial entities,

- open sharing and reuse of the tools and instruments developed in the research (e.g., software);
- database rights;
- the commercialisation of research results;
- legal and ethical issues arising from the use of social media content in research.

A separate group of discussed issues was related to the policies and guidelines of the National Science Centre. In addition to the issues already discussed above, the participants were interested whether RDM costs were allowable costs that may be included in grant proposals and whether it would be possible to opt out of the obligation to share research data.

In addition, inquires related to research data repositories could be identified among the frequently asked questions. The training participants were most often interested in obtaining information on the financial terms of use of repositories, long-term preservation of research data (often indicating 10 years), interoperability with other systems, databases or e-infrastructures (usually in terms of the visibility, findability and discoverability of research data), as well as the technical solutions and conditions for depositing data, including persistent identifiers, embargoes, suggested file formats and maximum size of files and data sets. These questions concerned both services established as part of the Disciplinary Open Research Data Repositories project and other publicly available repositories (often without specifying which services were concerned). The comments most frequently made in the questionnaires concerned specific examples of research data management plans, case studies, and practical information on data management and data sharing. Participants also expressed interest in participating in similar trainings, especially discipline-specific ones. The question of whether the training sessions would broaden the knowledge of research data was usually answered by the participants in the affirmative, claiming that the training was important and useful.

5. Conclusions and summary

The most frequent comments made by the participants concerned specific practical examples. Possible actions that can be taken to address these needs include making data management plans publicly available and reusable for training purposes, using case studies of data management plans, open research data, and data documentation. In addition, future training activities should take into account discipline-specific conditions that shape managing and data sharing practices.

The first step in this direction has already been taken within the framework of the Disciplinary Open Research Data Repositories project. In 2021, disciplinary RDM training began as part of the project and in cooperation with the Professor Robert Zajonc Institute for Social Studies (ISS) and the Institute of Philosophy and Sociology of the Polish Academy of Sciences (IFiS PAN). RDM training for the social sciences took place in January and May 2021. The programme took

into account the presentation of the Social Data Repository (RDS) and discipline-specific conditions of data sharing, resulting, inter alia, from ethical principles, especially the obligation to protect privacy, maintain confidentiality, and obtain informed consent from potential research participants. The presentation of case studies from various disciplines may be part not only of training activities but also of broadly understood communication activities aimed at promoting open science. These activities may also include various programmes aiming at greater researcher engagement with research data. It is worth taking advantage of the guidelines and programmes implemented in foreign scientific institutions and described in the publication Engaging Researchers with Data Management: The Cookbook [18].

One of the most important challenges facing Polish scientific institutions in the field of research data management is the development and implementation of openness policies that take into account the research data produced and collected by the scientific staff. This concerns both the adoption of appropriate documents to regulate this issue and comprehensive actions in different areas and at different levels to support the implementation of recommendations and obligations. This issue is also related to institutional conditions that must be taken into account by researchers when working with data, e.g. practical issues related to services and solutions for the secure storage of data, as well as legal issues related to the rules and principles of intellectual property management that are in force in the scientific institution. Communication on the available solutions, as well as training and consultations, constitute an important element of policy implementation, while feedback from the training participants may form the basis for developing an offer of training and planning other types of activities supporting researchers in research data management and sharing.

References

[1] *Documents available on the sub-site of the National Science Centre - Open Science* 2021, https://ncn.gov.pl/finansowanie-nauki/otwarta-nauka

[2] cOAlition S 2021 *Plan S Principles*, https://www.coalition-s.org/plan_s_principles/

[3] *Policy of the National Science Centre on Open Access to Publications* 2021, https://ncn.gov.pl/sites/default/files/pliki/zarzadzenia-dyrektora/zarzadzenieDyr-38_2020_ang.pdf

[4] *The Open Data and Re-use of Public Sector Information Act* 2021, https://dziennikustaw.gov.pl/D2021000164101.pdf

[5] 2021 *Data Opening Programme 2021 – 2027 Appendix to Resolution No. 28 of the Council of Ministers*, https://monitorpolski.gov.pl/M2021000029001.pdf 73

[6] *Regulation (EU) 2021/695 of the European Parliament and of the Council of 28 April 2021, establishing Horizon Europe – the Framework Programme for Research and Innovation, laying down its rules for participation and dissemination, and repealing Regulations (EU) No 1290/2013 and (EU) No 1291/2013* 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32021R0695from=EN

[7] Science Europe 2021 *Research Data Management*, https://www.scienceeurope.org/our-priorities/research-data/research-data-management/

[8] Science Europe 2021 *Broad Uptake of Science Europe Research Data Management Guide*, https://www.scienceeurope.org/news/broad-uptake-of-rdm-guide/

[9] *Letter from the NCN Director on the management of scientific data in projects* 2021, *National Science Centre*, https://ncn.gov.pl/sites/default/files/pliki/2019_04_03_pismo_dyrektora_NCN_zarzadzanie_danymi_naukowymi.pdf
[10] LERU 2021 *Sorbonne Declaration on research data rights*, https://www.leru.org/news/data-summit-in-paris
[11] Science Europe 2021 *Implementing Research Data Management Policies Across Europe – Experiences From Science Europe Member Organisations*, https://www.scienceeurope.org/media/jikjlb2g/se_rdm_best_practices.pdf 20
[12] *Account from the International Conference "Open Research Data: Implications for Science and Society"* 2015, *Open Science Platform*
[13] Fenrich W, Siewicz K and Szprot J 2016 *Towards Open Research Data in Poland*, http://pon.edu.pl/nasze-publikacje?pubid=19
[14] Fenrich W and Gruenpeter N 2020 *Domain Repository of Open Research Data*, *EBIB Bulletin [online]*, http://open.ebib.pl/ojs/index.php/ebib/article/view/698, **5** (194)
[15] *DMP Online* 2021, https://dmponline.dcc.ac.uk/
[16] https://www.ncn.gov.pl/aktualnosci/2019-05-07-posiedzenie-rady-kwiecien-2019
[17] *Research Data Management Plan - most frequently asked questions* 2020, *National Science Centre*, https://ncn.gov.pl/aktualnosci/2020-03-06-plan-zarzadzania-danymi-pytania
[18] Clare C, Cruz M, Papadopoulou E, Savage J, Teperek M, Wang Y, Witkowska I and Yeomans J *Engaging Researchers with Data Management: The Cookbook*, Open Book Publishers, Cambridge, doi: 10.11647/OBP.0185

Natalia Gruenpeter works at the Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw (ICM). Her responsibilities include communication, dissemination and training activities at the Open Science Platform, an initiative run by ICM.