

Promoting Common Understanding on Research Data Management using Rubric

Takaaki AOKI
Academic Center for Computing
and Media Studies
Kyoto University
Kyoto, Japan
ORCID: 0000-0002-5926-4903

Shoji KAJITA
Academic Center for Computing
and Media Studies
Kyoto University
Kyoto, Japan
ORCID: 0000-0002-6615-2176

Tamaki MOTOKI
Academic Center for Computing
and Media Studies
Kyoto University
Kyoto, Japan
ORCID: 0000-0003-4219-3122

Toshihiko IYEMORI
Academic Center for Computing
and Media Studies
Kyoto University
Kyoto, Japan
ORCID: 0000-0002-3648-0697

Tomoko KAWAGUCHI
Academic Center for Computing
and Media Studies
Kyoto University
Kyoto, Japan
ORCID: 0000-0003-1927-6931

Abstract — Research data management (RDM) is the activity to describe and practice what kind of data is to be used/obtained/generated, and how that data is analyzed/saved/shared or published from the plan to completion of the research project. The RDM consists ambiguous concept so that the understanding of RDM is diverse according to each researcher. In order to foster a common recognition of RDM in a Japanese university, a workshop that researchers utilize a rubric to evaluate and review their attitudes to RDM was held. In this paper, we report the process of the development of RDM rubric and the workshop.

Keywords—research data management, open science, fair principle, capability maturity model, rubric

I. INTRODUCTION

Research data management (RDM) is the activity to describe and practice what kind of data will be used / acquired / generated, and how is such data analyzed / stored / shared / release. In recent years, the importance of RDM has been addressed in various aspects, such as promoting open science and strengthening research fairness.

In current European countries, open science has been recognized as the method to explore new scientific field and contribute to citizen science by sharing and disclosing research data, especially generated through the research activities by public funds. In order to promote the research data open, a comprehensive research data management methods such as acquisition, storage, and sharing of research data, has been defined. For example, in United Kingdom, "Concordat on Open Research Data" [1] was issued on 2016, which states the significance of open research data and the action guidelines for each stakeholder. In addition, major research funders including Research Councils UK request the researchers to prepare data management plan (DMP) attached to the research proposal [2],

which encourages the appropriate management of data is performed actively.

In Japan, the interest in research data management started from the viewpoint of research integrity around 2013, when the Ministry of Education, Culture, Sports, Science and Technology Japan updated "Guidelines for Responding to Misconduct in Research" [3] in response to frequent incident of research misconducts. Following these guidelines, higher education institutions such as universities have established policies for storing research data [4]. On the other hand, the rules oriented to open science promotion are also developing. some Japanese funding agencies started to mandate their research projects to formulate DMPs [5, 6, 7]. The Cabinet Office issued "Guideline for developing data policy at National Research and Development Agency" [8].

Contrary to the development of policy and rules from the top down, many researchers have doubt about the effectiveness and practical issues in a deployment of these rules, that is recognized as the gap. According to the survey on data management, preservation and publication by Kurata et al. in 2016 [9], in many institutions, there are rules and guidelines to preserve research data at the request of [3]. However, it is suggested that the concept of openness of research data and importance of RDM to support open research data are not fully understood or accepted by the university researchers and the research support staff.

Considering RDM from the researcher side, it is a natural for research activity to handle research data efficiently and to prevent data loss and falsification regardless of intentional or accidental. In addition, there are many cases regardless of the research field, in which research data are organized in database and open for public. Moreover, there is a consensus to share the research data in the entire academic field, such as life sciences, planet earth sciences, etc. In other words, it can be said that,

although researchers have already implemented their own RDM in some form, there is a very large divergence in the method of RDM implementation depending on researchers and academic fields. This also suggests that it is difficult to implement a specific procedures according to the top-down policy.

II. RUBRIC ON RESEARCH DATA MANAGEMENT

Ideally, the RDM procedure implemented individually by researchers at present, should progressively transformed to one advocated by a unified policy. To achieve it, it is essential to unify the definition and term related to the RDM and to develop research data handling skills. The materials for RDM literacy education for this purpose are actively being developed. Also in Japan, “RDM Training Tool” [10] and the online training course “Research data management in open science era” were published in 2017, and gathered many users [11].

In addition to RDM education utilizing training materials, it is necessary to have a method to objectively evaluate the understating and implementation to RDM of individual researcher or organization. The capability maturity model (CMM) is major technique to evaluate the achievement and formulate the strategy of organizations’ software development process. A rubric is a tabulated list of the level of achievement for each process areas related to the goal of the organization. The idea of CMM has been introduced to scientific data and research data management by several groups. For example, the rubric proposed by Qin et al. [12, 13] offers 5 large sections (general, acquisition, description, dissemination, preservation) with 61 subitems. Australian National Data Service proposes other model which consists of 5 elements (policies and procedure, IT infrastructure, support services, managing metadata, managing research data) to be assessed [14]. Both two models are designed for institutes or large-scale data service providers.

TABLE 1.THE “SUPPORT YOUR DATA” RDM RUBRIC [15]

	Ad Hoc	One-Time	Active and Informative	Optimized for Re-Use
Planning your project	I don’t follow a consistent approach for keeping my data organized, so it often takes time to find things.	I have an approach for organizing my data, but I only put it into action after my project is complete.	I have an approach for organizing my data that I implement prospectively, but it not necessarily standardized.	I organize my data so that others can navigate, understand, and use it without me being present.
Organizing your data	I don’t follow a consistent approach for keeping my data organized, so it often takes time to find things.	I have an approach for organizing my data, but I only put it into action after my project is complete.	I have an approach for organizing my data that I implement prospectively, but it not necessarily standardized.	I organize my data so that others can navigate, understand, and use it without me being present.
Saving and backing up your data	I decide what data is important while I am working on it and typically save it in a single location.	I know what data needs to be saved and I back it up after I’m done working on it to reduce the risk of loss.	I have a system for regularly saving important data while I am working on it. I have multiple backups.	I save my data in a manner and location designed maximize opportunities for re-use by myself and others.
Getting your data ready for analysis	I don’t have a standardized or well documented process for preparing my data for analysis.	I have thought about how I will need to prepare my data, but I handle each case in a different manner.	My process for preparing data is standardized and well documented.	I prepare my data in such a way as to facilitate use by both myself and others in the future.
Analyzing your data and handling the outputs	I often have to redo my analyses or examine their products to determine what procedures or parameters were applied.	After I finish my analysis, I document the specific parameters, procedures, and protocols applied.	I regularly document the specifics of both my analysis workflow and decision making process while I am analyzing my data.	I have ensured that the specifics of my analysis workflow and decision making process can be understood and put into action by others.
Sharing and publishing your data	I share the results of my research, but generally I do not share the underlying data.	I share my data only when I’m required to do so or in response to direct requests from other researchers.	I regularly share the data that underlies my results and conclusions in a form that enables use by others.	Because of my excellent data management practices, I am able to efficiently share my data whenever I need to with whomever I need to.

Contrary, the CMM and the rubric designed for small laboratory or individual researcher were proposed by J. Borghi et al. at California Digital Library through the project “Support Your Data” [15]. The RDM rubric has the following six evaluation criteria: This is in line with the research data life cycle phase.

1. Planning your project

2. Organizing your data

3. Saving and backing up your data

4. Getting your data ready for analysis

5. Analyzing your data and handling the outputs

6. Sharing and publishing your data

At each stage, four degrees of maturity, “ad-hoc”, “one-time”, “active and informative”, and “optimized for re-use”, are set. The contents of the rubric are intended to be in compliance with the FAIR principle [16] and its literary style is designed not for experts in IT nor data management but for all researchers regardless of academic fields and generations.

III. ORGANISING WORKSHOP FOR UTILIZING RDM RUBRIC

In order to try to extract the differences in individual thinking for RDM and to identify the common issues for achieving more mature RDM, a workshop utilizing RDM rubric by CDL was organized by the Kyoto University Academic Data Innovation Unit [17]. The unit was established in 2017 with the aim for interdisciplinary study and development of research management environment at the university.

In this workshop, ten people from inside and outside of Kyoto University participated in the workshop. They have certain background knowledge in research data management, but they did not know the concept CMM and rubric for RDM. The participants self-assessed the individual attitude to RDM by giving a brief review of “Why does it fit into the level of maturity” for each evaluation item of the rubric. Instead of writing in the separated form, we adopted a method that the participants wrote simultaneously on the common Google Document. By referring to the answers of others, the difference in interpretation between the evaluation criteria and the maturity level was eliminated, which was succeeded to draw proper answers without any confusion from all participants. It took about 30 minutes to complete the answer to the six evaluation criteria, and then 30 minutes for cross-reviewing to the results.

Prior to answering to the rubric, it was necessary to clarify the relationship between research data and each person. In this group working, The participants were classified into the following roles.

Solo: The research is conducted by almost one person (4 persons)

PI: Leader in the laboratory, mainly acts as a project manager (3 persons)

Arc: As an archivist or a content creator, who works to store and process research data generated by others (2 person)

Repo: is responsible for managing data repository and supports other researchers to register and utilize research data.(1 person)

The discussion by cross-reviewing the answers by the participants are summarized as following.

A. Planning your project

As the practice of creating DMP in advance has not yet been established in Japan, many respondents answered that it corresponds to Ad-hoc. Among them, the answer from Repo was “optimized and re-use”. It is evaluated that the data acceptance criteria for the data repository is already prepared and presented, means the data management is mandated to the relevant data repository poster. From the viewpoint of intellectual property management, there were also opinions that it would be necessary to create DMP when concluding a joint research contract.

B. Organizing your data

As for naming files and folders, Solo has established rules for their own use and other roles has stated the common rules for groups, which results in the answer of “Active and Informative”. However, it is difficult to adjust file naming conventions when working between different groups, and some people have asked for the establishment of more acceptable rules. The Arc and Repo roles pointed out that, at the metadata level, the method of describing and organizing data has been established as rules or manners.

C. Saving and backing up your data

From any of the roles, “One-Time” or “Active and Informative” was answered. They have obtained backups in some way, such as archiving storage, cloud storage, source code repository, etc., but there is little use of systematic and secure methods guaranteed or supported by their institutes. Also, it is difficult to handle large volume data which cannot be backed up practically like simulation data. It was pointed out that the need for a technology to sort and discard data as data became bloated.

D. Getting your data ready for analysis

The answer was almost the same as “Organizing your data.” The data preparation by Solo tend to be “Ad-hoc” because he usually describe the data structure as she/he can understand by herself/himself, but sometimes misses the minor change of data format other information. In other roles, rules have been defined for efficient use of research data on the premise of joint work or third party provision, so that it was evaluated as optimized and re-use.

E. Analyzing your data and handling the outputs

In any of the roles, the analysis program is maintained in consideration of data re-use and repeatability. Preservation of analysis program and parameter list works to remember “what was done.” However, the methodology for recording the decision-making process, “why did it,” is not established, so that it was often evaluated as “one time”.

F. Sharing and publishing your data

With the development of institutional repositories and the progress of open access by publishers, documents such as papers and reports can be opened by an individual effort. However, the activity for open research data is evaluated as insufficient by Solo and PI researchers. This is because that, there is no appropriate IT infrastructure and it might be a problem of the qualification of their research data. The other problems was pointed that, in the research dealing with personal information, such as onymous survey and interviewing, preceding agreement with the data provider should be mandated and prepared at the stage of “Planning your project.”

IV. SUMMARY

A workshop was conducted using the RDM rubric proposed by CDL to objectively evaluate the RDM implementation considered by individual researchers. Although the participants at this time had preliminary knowledge about open science and RDM, the workshop was accepted as a fresh experience to verbalize the degree of RDM practice and to evaluate them mutually. In particular, because the method on this workshop initiates the thought for standard RDM from the bottom-up, it is

expected that this trial will fill in the gap with top-down determined RDM policy.

The authors expect that by organizing similar workshops repeatedly, the examples of concrete implementation at each maturity level will be presented and a common understanding of RDM at the field level will be fostered. The practices by bottom-up activities is essential for clarifying organizational responses to facilitate more advanced RDM environment including IT service and supporting human resources.

ACKNOWLEDGEMENT

The authors appreciate to Dr. Stephanie Sims of the California Digital Library and the participants to “the first Kyoto University Data Management Workshop”, for the collaboration to utilize “Support your data” rubric. We also acknowledge to Center for Interdisciplinary Studies and Education of Kyoto University for supporting the activity of the Kyoto University Academic Data Innovation Unit and the operation of the workshop.

REFERENCES

- [1] “Concordat on Open Research Data,” 2016. “Concordat on Open Research Data,” 2016 <https://www.ukri.org/files/legacy/documents/concordat-on-open-research-data-pdf/>
- [2] “Guidance on best practice in the management of research data”, 2015 <https://www.ukri.org/files/legacy/documents/rcuk-common-principles-on-data-policy-pdf/>
- [3] Ministry of Education, Culture, Sports, Science and Technology, “Guidelines for Responding to Misconduct in Research”, 2014, http://www.mext.go.jp/a_menu/jinzai/fusei/1359618.htm
- [4] T. Aoki, S. Kajita, H. Akasaka, and H. Takeda, “Development and Deployment of Research Data Preservation Policy at a Japanese Research University in 2016,” in 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 2017, pp. 120–123.
- [5] Japan Science and Technology Agency, “JST Policy on Open Access to Research Publications and Research Data Management.” 2017 [Online]. Available: <https://www.jst.go.jp/EN/about/openscience/index.html>.
- [6] Japan Agency for Medical Research and Development, “データマネジメントプランの提出の義務化について [The obligations of the submission of the data management plan]” 2018. [Online]. Available: <https://www.amed.go.jp/koubo/datamanagement.html>.
- [7] New Energy and Industrial Technology Development Organization, “NEDO プロジェクトにおけるデータマネジメントについて [Data Management Policy of the Projects Awarded by NEDO]”, 2019 [Online]. Available https://www.nedo.go.jp/jyohoukoukai/other_CA_00003.html
- [8] Japan Cabinet Office, “国立研究開発法人におけるデータポリシー策定について [Guideline for developing data policy at National Research and Development Agency]”, 2018, [Online]. Available: <https://www8.cao.go.jp/cstp/stsonota/datapolicy/datapolicy.html>
- [9] K. Kurata, M. Matsubayashi, and M. Takeda, “Research data management in Japanese universities and research institutions”, Joho Kanri, vol. 60, np. 2, pp. 119-127. Feb 2017.
- [10] Japan consortium for open access repository, “RDM Training Tools” 2017. [Online]. Available: <http://id.nii.ac.jp/1458/00000023/>
- [11] M. Furukawa, K. Ojio, and K. Yamaji, “Development and Analysis of RDM Training Online Course” 情報教育シンポジウム論文集, pp. 84-89, 2018, Available: <http://id.nii.ac.jp/1001/00190691/>.
- [12] J. Qin, K. Crowston, and A. Kirkland, “Pursuing Best Performance in Research Data Management by Using the Capability Maturity Model and Rubrics,” J. eScience Librariansh., vol. 6, no. 2, p. e1113, 2017.
- [13] “A Capability Maturity Model for Research Data Management.” [Online]. Available: <http://rdm.ischool.syr.edu/xwiki/bin/view/CMM+for+RDM/WebHome>.
- [14] “Creating a data management framework.” [Online]. Available: <https://www.andis.org.au/guides/creating-a-data-management-framework>.
- [15] J. Borghi, S. Abrams, D. Lowenberg, S. Simms, and J. Chodacki, “Support Your Data: A Research Data Management Guide for Researchers,” Res. Ideas Outcomes, vol. 4, p. e26439, May 2018.
- [16] M. D. Wilkinson, et al., “The FAIR Guiding Principles for scientific data management and stewardship,” Sci. Data, vol. 3, p. 160018, Mar. 2016.
- [17] “Kyoto University Academic Data Innovation Unit”, <http://www.cpiir.kyoto-u.ac.jp/about/acdi/>