Development and Analysis of Online RDM Training Course

Masako Furukawa
National Institute of Informatics
Tokyo, Japan
furukawa@nii.ac.jp

Koichi Ojiro
National Institute of Informatics
Tokyo, Japan
ojiro@nii.ac.jp

Kazutsuna Yamaji
National Institute of Informatics
Tokyo, Japan
yamaji@nii.ac.jp

***Abstract*— Research data management (RDM) is important to support research data sharing in open science, as well as to enhance the transparency of the research process. However, there are still very few staff trained in research data management in Japan. Therefore, the Open Access Repository Promotion Association, known as JPCOAR, organized a research data task force and developed a training tool to teach research data management skills. National Institute of Informatics (NII) collaborated with the JPCOAR to launch an online course "Research data management in the open science era" on November 15, 2017. In this paper, we discuss development and analysis of the online RDM training course.**

Keywords—MOOC, Research Data Management

I. INTRODUCTION

Research data management (RDM) is a term that refers to a series of tasks related to the organization, structuring, preservation, sharing, publishing, and reuse of data in research projects. RDM is important to support research data sharing in open science, as well as to enhance the transparency of the research process.

Japan participated in a joint declaration on open science at the G8 in 2013, and signing this declaration triggered various national policies [1]. For example, with the aim of stimulating innovation, publicly funded researches are required to provide easier and wider access to their research results. As for research fairness, in 2014, the Ministry of Education, Culture, Sports, Science and Technology provided guidelines for dealing with research data, and requested that universities formulate rules for this purpose [2]. In terms of data preservation, the Science Council of Japan outlined guidelines for preserving research data for 10 years [3]. Based on this guideline, the University of Tokyo has established rules for data storage and preservation.

In the case of the UK, there are various research funding agencies. These agencies place several obligations on research data management [4]. Such trends will also spread to Japan. There are several Japanese funding agencies that specify policies on research data management. For example, The Japan Agency for Medical Research and Development (AMED) requires a research management plan when applicants apply for a grant [5]. Japan Science & Technology Agency (JST) has also set out a basic data policy [6], and several funding agencies also require applicants to submit a data management plan. The Ministry of Economy, Trade and Industry has a fund called MED. To apply for this fund, applicants are required to provide a data management plan.

On the other hand, the publisher Springer Nature [7] applies four different research data policies concerning its journal publications. The University of Edinburgh [8] specifies a policy on management, sharing, preservation, and disclosure of research data. Although the policy of many Japanese universities only regulate the preservation of data, the policy of the University of Edinburgh stresses, not only preservation, but also the sharing of data. This trend is likely to spread across Japanese universities.

A summary of deliberations by the Ministry of Education, Culture, Sports, Science and Technology states that: one of the roles expected of universities is to support researchers by creating a data management schema, where technical staff, URA(University Research Administrator)s, and university library staff play an important role. In order to put this into practice, appropriate support is necessary. Many countries offer teaching materials that show how to manage such data [9-13].

However, there are still very few staff trained in research data management in Japan. Therefore, the National Institute of Informatics collaborated with the JPCOAR to launch an online course ("Research data management in the open science era") on November 15, 2017. In this paper, we discuss the development and analysis of this online RDM training course.

II. DEVELOPMENT OF THE ONLINE RDM TRAINING COURSE

The RDM Training Course is based on the teaching materials created by JPCOAR [14]. JPCOAR was established in 2016 by integrating three communities: the Digital Repository Federation (DRF), well-known to librarians; the promotional community JAIRO Cloud; and the Institutional Repository. JPCOAR is now a very large community of institutional repositories, with more than 500 institutional participants, mainly university libraries. One of JPCOAR's activities is to develop leading-edge functions in institutional repositories. Among them, the Research Data Task Force has been developing teaching materials on the fundamentals of research data management.

The aim of the teaching material is to provide learners with a basic knowledge of research data management, and give

them a foundation on which to build research data management services. The course is divided into 7 chapters, and each chapter consists of slides, lecture notes and confirmation tests. The teaching materials aim to support the life cycle of research data, i.e. Data Creation, Data Processing, Data Analysis, Data Preservation, Data Access, and Data Reuse. This teaching material was released with CC-BY license on JPCOAR's website on June 6, 2017.

The configuration of the teaching material is as follows.

Chapter 1: The Introduction, which explains the background to RDM requirements, and gives a definition of research data and RDM.

Chapter 2: The definition of Data Management Policy (DMP), explaining trends and how to formulate actual DMP. This chapter also introduces tools to support DMP.

Chapter 3: The preservation and sharing of research data, explained in the context of security and the utilization of data repositories for sharing/reuse.

Chapter 4: Systematic management techniques for reusing data, including metadata schemas.

Chapter 5: Legal and Ethical Issues. Copyright licensing and policies on research fraud.

Chapter 6: Policies on research fairness and the handling of research data.

Chapter 7: Designing RDM services, human resources to support RDM, and basic systems of research data preservation and disclosure.

Based on these teaching materials, the National Institute of Informatics collaborated with the JPCOAR to launch an online course ("Research data management in the open science era") on November 15, 2017. The online course was built on "gacco", which is one of the major MOOC platforms in Japan.

The large-scale public online courses known as MOOC provide university-level educational opportunities free of charge. It initially spread in the United States in 2012. In Japan, JMOOC was established in 2013 [15]. JMOOC currently has four platforms and about 250,000 registrants. The number of learners has exceeded 610,000. The general course at JMOOC is composed of four weeks of study. If students submit tasks and comprehensive assignments each week and satisfy the completion requirements, a certificate of completion is issued.

We reorganized the seven chapters of the RDM teaching materials mentioned above into a four-week course for MOOC (Fig.1). In the first week, students learn about the importance of research data management. In the second week, the focus is about sharing and documenting research data. The third week deals with metadata and legal/ethical issues. In the fourth week, we examine how to design services to manage research data at universities and research institutions. Each week of the MOOC course consists of 4 to 5 videos, download materials and a confirmation test (10 questionnaires). In addition, there are discussion boards and optional questionnaires. We issued a certificate of completion for learners achieving 70% of the score for each week's confirmation test.

Fig. 1. Online RDM training course. Top page

III. Platform for Learning Data Analysis

Every time a student uses a module of the courseware, a new learning-record entry is added to the learning management system. The learning records include the grades for online assignments, reading time, the total number of login times, the total number of online discussions and so on. Learning analytics is the technology used to collect and analyze educational data. Learning analytics is useful for evaluating and improving the courses, and enables adaptive support for learners [16].

Regarding sharing, LearnSphere [17] was developed under the auspices of the National Science Foundation (NSF). LearnSphere is an infrastructure for learning analysis. By simple drag-and-drop operations, processing the designation of storage location, selection of statistical model and visualization of learning data can be performed. However, responding to international standardization and interoperability of tools and data are future subjects.

For the sharing and utilization of educational resources, the National Institute of Informatics has provided various services such as the development of teaching materials and a learning environment as an inter-university research institute. The academic information network (SINET), managed by NII, has been used as an academic information infrastructure for universities and research institutes throughout Japan. In addition to these functions, we have developed the Learning Analytics platform.

The architecture of the Learning Analytics platform is shown in Fig.2. The learning records are transferred to the Log Store server and are used for the analysis. The learning records from multiple servers are gathered to one log store server. The base of the server is Learning Locker, which is an open source log store server. The mongoDB is used as a database server.

In our system, the learning log data extracted from the LMS are translated to the xAPI [18]. xAPI is one of the standards related to learning log analysis. xAPI describes learning log data with JSON format, and enables applications to share log data. 90 kinds of events, such as “logged in” and “viewed”, are translated to the xAPI statements.

To protect personal information, personally identifiable data such as user name, user ID, IP address are anonymized using SHA-2 hash. If other personally identifiable data are hashed in the same way, we can combine learning records from multiple data sources.

The results of basic analysis, such as simple statistical values, are shown on the dashboard interface (Fig.3). The detailed analysis of learning records is also available on the Web browser. The analysis can be carried out using statistical analysis software R. The advantage of our system is that we can share the analysis program on the dashboard, and can use the dashboard as a basis for research and improvement of the courses. The developed system can be used as a platform for the accumulation, analysis and sharing of educational contents, together with functioning as the institutional repository of NII.

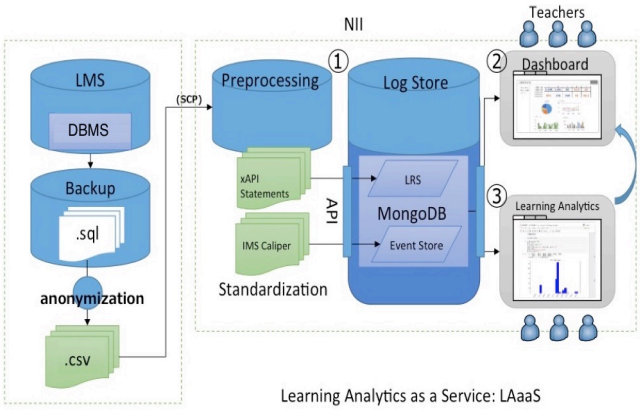

Fig. 2. Platform for learning data analysis.

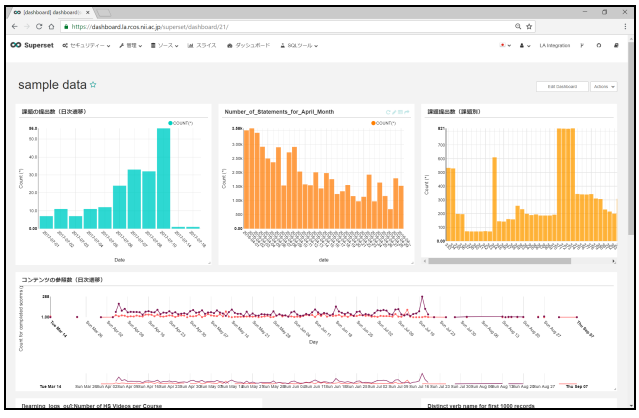

Fig. 3. Dashbord interface.

IV. Analysis of the Online RDM Training Course

The MOOC course, “Study data management in the open science era”, ran from November 15, 2017 until January 15, 2018. The number of participants in the course was 2305. The number of registrants in the course exceeded expectations for such a highly specialised program. The completion rate of 25% was very high compared to the gacco’s average completion rate of 15%.

We carried out a questionnaire before the course. Although responses to this questionnaire were not essential, there were 770 respondents. As for the gender, 61% of the participants were male and 39% were female. Fig.4 shows the birth year of the participants. We can see that there are many participants in their 40s. As for the familiarity with research data management, about 80% of the respondents said that they already knew a certain amount about research data management (Fig.5). 616 respondents answered that they were doing full-time jobs. About 50% of these worked at a university or research institute, and about 30% worked in the library system (Fig.6).

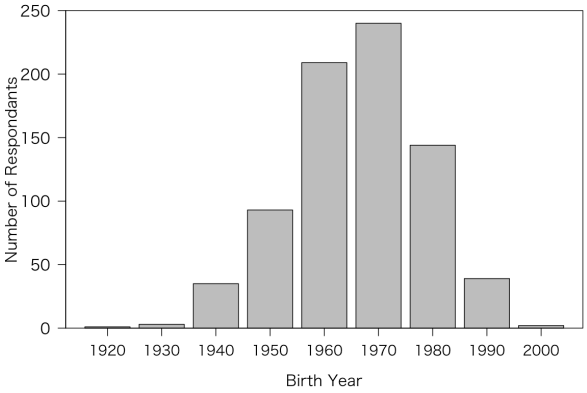

Fig. 4. Birth year of participants.

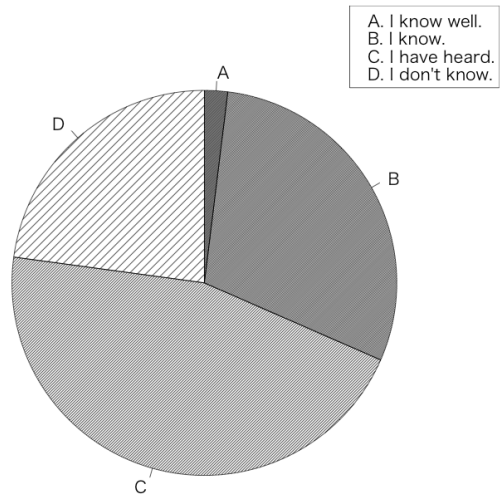

Fig. 5. Familiarity with Research Data Management.

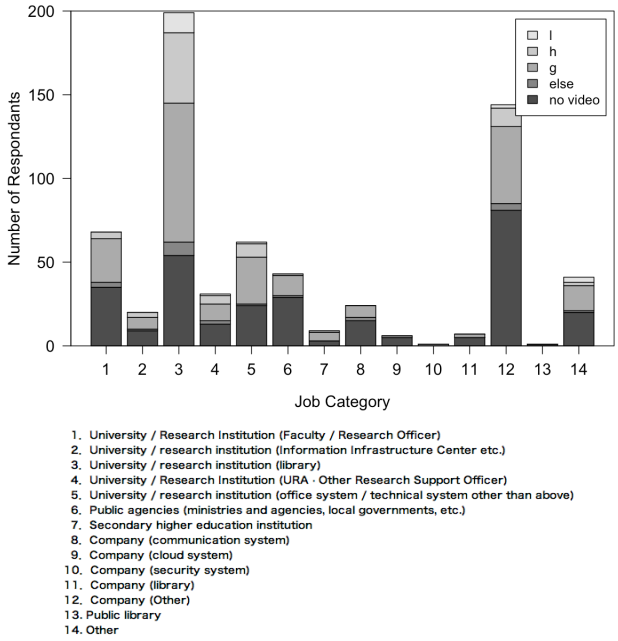

Fig. 6. Job category of respondents.

The main materials of this course are videos. Therefore, we analyzed whether there was a difference in video access by learners. The analysis was carried out for learners who watched at least one video. The number of times a video was accessed each week was collected over 4 weeks, and we used these 4 dimensional data as features of learners. We used k-means method to cluster these 4 dimensional data. In the case of video access errors, the number of times a video is recorded as being accessed can sometimes become enormously large. By choosing a cluster number of 15, data are clustered into small clusters, and we can reduce the effects of such clusters.

Fig.7 shows the number of Passes/Fails for each cluster. Almost all the learners are in 3 clusters: cluster g, h and l. In the 3 clusters, cluster h and l have the highest completion rates.

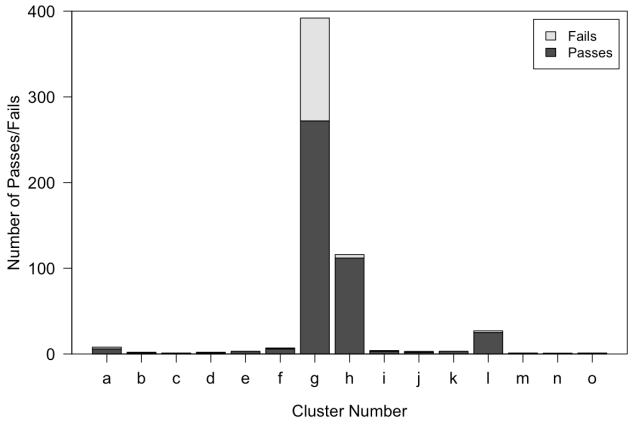

Fig. 7. Number of Passes/Fails for each cluster.

These 2 clusters correspond to learners who accessed videos a large number of times (Fig.8). The number of each cluster according to job category is shown in Fig.6. Job category 3, which corresponds to librarians, contains many of the cluster h and l. This means that many of the librarians who took this course watched videos often.

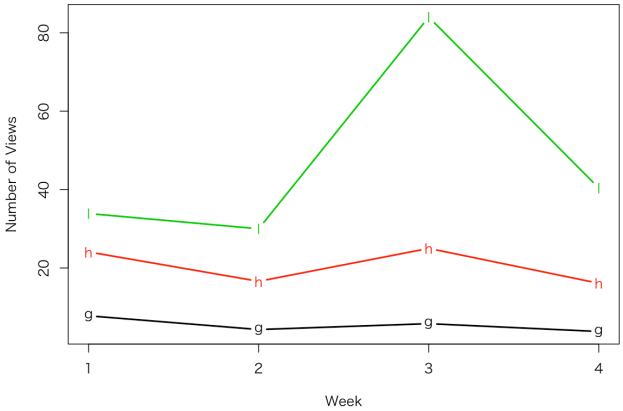

Fig. 8. Number of Video Views for each cluster.

345 participants responded to the questionnaire after the course. The distribution of gender, age group, and job category were very similar to the initial questionnaire. About 90% of learners responded saying that the course had met their expectations (Fig.9).

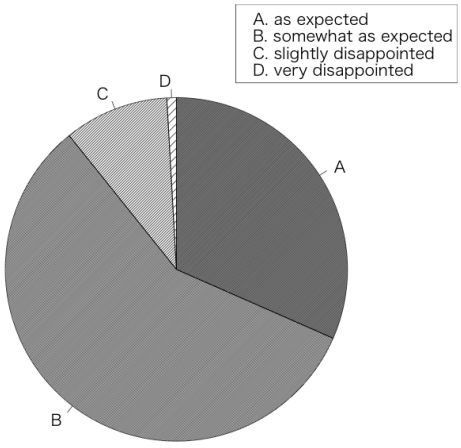

Fig. 9. Expectations of the course.

Learners were also asked to elaborate on their answers. Learners who answered “as expected” or “somewhat as expected” described as follows.

“I was able to systematically learn the outlines of research data management.”

“What I need to do in the future became clear.”

“A lot of domestic and overseas examples were given. I found this very helpful.”

Learners who answered “slightly disappointed” or “very disappointed” described as follows.

“Although the contents are good, I was disappointed that lecturers did not appear much in the video.”

“I could easily answer the comprehension quiz using existing knowledge and references without watching the video at all.”

About 90% of learners replied that they would like to recommend this course to others (Fig.10). Learners were also asked to elaborate on their answers. Learners who answered “I highly recommend it” or “I would recommend it slightly” described as follows.

“I could learn from the basics; I think speed and length of the videos are appropriate.”

Learners who answered “I do not recommend it much” or “I do not recommend it” described as follows.

“Many parts are difficult to understand without experience of research at university.”

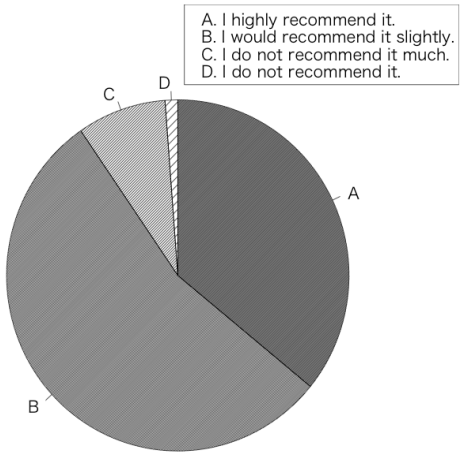

Fig. 10. Recommendation of the course.

V. Conclusions

In this paper, we discussed the development and analysis of an online RDM training course. In summary, many of the participants are librarians, there was a high level of satisfaction with the course, and participants were also active in disseminating information about the course. It can be said that we achieved our original objective.

We hope that librarians will use this teaching material in the following ways: First, for self-study. Second, to incorporate research data management into computer literacy education programs. We hope computer literacy programs enable graduate students and young researchers to learn about research data management. Third, to develop on-campus training for researchers, URAs, or technical staffs.

Our next step is to develop new teaching materials for support staff based on the results of the MOOC course. The teaching materials are designed to show what services to learn and how to put them into practice. Preparation of the teaching materials will be launched in the financial year of 2018.

References

[1] A joint statement by the G8 Science Ministers on making research data open(G8), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf

[2] 「Guidelines for Responding to Misconduct in Research」（Adopted August 26, 2014 by Ministry of Education, Culture, Sports, Science and Technology （MEXT）） , http://www.mext.go.jp/a_menu/jinzai/fusei/__icsFiles/afieldfile/2015/07/13/1359618_01.pdf

[3] Science Council of Japan (SCJ), "Recommendations Concerning an Approach to Open Science that Will Contributes to Open Innovation", http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-23-t230-en.pdf

[4] The Digital Curation Centre, http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

[5] Data sharing Policy for the Realization of Genomic Medicine, https://www.amed.go.jp/content/000017356.pdf

[6] JST Policy on Open Access to Research Publications and Research Data Management, https://www.jst.go.jp/EN/about/openscience/policy_openscience_en.pdf

[7] Research Data Policy Types, https://www.springernature.com/gp/authors/research-data-policy/data-policy-types/12327096

[8] Research Data Management Policy, https://www.ed.ac.uk/information-services/about/policies-and-regulations/research-data-policy

[9] MANTRA, https://mantra.edina.ac.uk/

[10] FOSTER, https://www.fosteropenscience.eu/

[11] RDMRose, http://rdmrose.group.shef.ac.uk/

[12] figshare, https://knowledge.figshare.com/open-data/about-rdm

[13] zenodo, https://zenodo.org/

[14] JPCOAR, https://jpcoar.repo.nii.ac.jp/

[15] JMOOC, https://www.jmooc.jp/en/

[16] New Media Consortium. Learning Analytics and Adaptive Learning. NMC Horizon Report: 2016 Higher Education Edition, 2016, p.38-39.

[17] LearnSphere, http://learnsphere.org/

[18] ADL, “Experience API Specification”, https://github.com/adlnet/xAPI-Spec/blob/master/xAPI.md