Bausteine Forschungsdatenmanagement
Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern

Designing and Implementing Practicable Data Management Plans in Large-Scale Projects

Soo-Yon Kim[i] Steffen Hillemacher[ii] Stefan Decker[iii]
Bernhard Rumpe[iv] Sandra Geisler[v]

2023

Zitiervorschlag

Kim, Soo-Yon, Steffen Hillemacher, Stefan Decker, Bernhard Rumpe, Sandra Geisler. 2023. Designing and Implementing Practicable Data Management Plans in Large-Scale Projects. *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern* Nr. 3/2023: S. 2-12. DOI: 10.17192/bfdm.2023.3.8571.

[i]Information Systems & Databases, RWTH Aachen University. ORCID: 0000-0001-5975-0031
[ii]Software Engineering, RWTH Aachen University. ORCID: 0000-0002-6819-9031
[iii]Information Systems & Databases, RWTH Aachen University. ORCID: 0000-0001-6324-7164
[iv]Software Engineering, RWTH Aachen University. ORCID: 0000-0002-2147-1966
[v]Data Stream Management & Analysis, RWTH Aachen University. ORCID: 0000-0002-8970-6282

Abstract

Creating a data management plan serves as both a suitable entry point as well as an accompanying guide to research data management for researchers. While the practice is becoming more common, it is still uncharted territory for most researchers. A comprehensive guidance as well as a facilitating environment are essential to enable researchers to integrate this practice into their everyday work. So far, little research has been directed towards the implications for the work of data stewards in this context, with hardly any recommendations on how they can provide such an enabling guidance and environment. We explored the implications in the Cluster of Excellence "Internet of Production" by conducting a series of interviews with researchers on the topic of research data management practices. Five major requirements emerged for the practicable design and implementation of a data management plan template in large-scale projects. On their basis, a customized template was created, whose project-wide implementation is currently being carried out.

1 Introduction

As the importance of data as a valuable resource increases in all areas of society, innovation in data-driven methods is being rapidly advanced in science as well, and research data management has become an important component of good scientific practice. As data management plans (DMPs) are built around the data life cycle and thus address all essential aspects of research data management (RDM),[1] creating a DMP poses a useful method for researchers to put extensive and structured thought into and apply measures to the management of their research data. There is already a variety of DMP templates in the form of questionnaires available, e.g., provided by funding organizations, such as in the Horizon Europe program,[2] or by research facilities, such as the RWTH Aachen University.[3] In theory, creating a DMP for a project is as simple as choosing a template and distributing it to the project researchers for them to answer its questions. Practically, having a DMP completed and ensuring the quality of the information content of the filled-in document demands several considerations, especially concerning the needs of the researchers who need to integrate working with DMPs into their daily research routine. With regard to the work of data stewards, little research exists on how to design and implement DMP templates in their projects in such a way that these challenges are tackled and adequate guidance and an enabling environment for researchers is provided. This may include recommendations on the

[1] Michener, William K. "Ten Simple Rules for Creating a Good Data Management Plan." *PLoS Computational Biology* 11, no. 10 (2015). https://doi.org/10.1371/journal.pcbi.1004525.

[2] European Commission, "Horizon 2020 FAIR Data Management Plan (DMP) template", accessed January 13, 2023, https://ec.europa.eu/research/participants/docs/h2020-fundingguide/cross-cutting-issues/open-access-data-management/data-management_en.htm#A1template.

[3] RWTH Aachen University, "Forschungsdatenmanagement von A – Z", accessed January 13, 2023, http://www.rwth-aachen.de/cms/root/Forschung/Forschungsdatenmanagement/~svkj/A-bis-Z.

choice of templates and tools as well as suggestions on communication structures and accompanying measures, basing them on the characteristics of the project environment. Our research aims at using the Cluster of Excellence "Internet of Production" as a use case to explore the requirements for the work of data stewards in DMP design and implementation at the example of a large-scale project and to implement our findings on it.

The rest of this paper is structured as follows. In Section 2, we present the use case in more detail and explain how the requirements for the project-specific DMP template were derived. Building on these, Section 3 shows how the DMP template is structured and its question flow was developed. In Section 4, its realization using a specific software tool is depicted and the concept of a project-wide roll-out of the DMP and its tooling is explained. Finally, Section 5 concludes this paper and provides an outlook on future work.

2 Deriving DMP Requirements

In order to retrace how we derived specific requirements for the DMP template, we will first discuss the setting of the use case, as it is important to understand the research project the template is designed and implemented for, especially regarding its organizational structure. Next, we will outline the aims we pursued for the DMP. Lastly, we introduce the method we used to derive the requirements and the results we found.

2.1 Use Case: The Internet of Production

Over the last decades, the role of data in production environments has gained considerable importance. The amount and variety of new data sources and tools, as well as the degree of connectivity between the domains along the entire production chain, have increased drastically. In this light, the research of the Cluster of Excellence "Internet of Production" (IoP) at the RWTH Aachen University revolves around enabling data-driven manufacturing in order to improve production processes and products by making the massive amounts of data available for cross-domain processing, analysis, and exchange.[4] Naturally, research in data-driven cross-domain engineering features a high degree of scientific collaboration and exchange of research data. Hence, the comprehensible documentation of research data in this increasingly complex environment is becoming both more challenging and relevant.

The IoP — a large-scale project. Around 200 researchers from more than 30 different institutes are involved, classifying the IoP as a large-scale project. The setting is

[4] RWTH Aachen University, "Exzellenzcluster Internet of Production", accessed January 13, 2023, http://www.iop.rwth-aachen.de.

characterized firstly by a high degree of interdisciplinarity, where disciplines are ranging from mechanical engineering to computer science to social sciences, with each discipline coming along with its own research culture; secondly, by constantly fluctuating staff, due to the sheer number of positions as well as the rather frequent personnel change which is typical in the scientific domain; and thirdly, by interorganizational collaboration, with the involved institutes having differing administration styles and infrastructures.

RDM in the IoP. The IoP employs two data stewards who are responsible for the implementation of RDM structures in the project. The team maintains regular contact with the project management, as well as the RDM team of the RWTH Aachen University and further partners such as the NFDI4Ing consortium.[5] An information portal and several project-wide assemblies per year serve as the main communication infrastructure. As for the data infrastructure, in addition to a GitLab instance which is provided project-wide, the researchers work with individual data management solutions.

2.2 Aims

We set the practicability of the DMP template as our major goal. The term "practicable" can be thought of as bridging the gap between what is ideal and the actual state of affairs, namely in terms of the DMP template, to provide a document in such a way that researchers will actually work with it and integrate it into their everyday research work. We can characterize practicability by three dimensions:

- Feasibility: Which constraints exist and which resources are necessary?
- Usability: What contributes to an efficient and easy use?
- Usefulness: Which of the researchers' interests are met?

To derive the requirements for achieving practicability, an analysis of the project conditions and input by the researchers was required. We therefore questioned the researchers.

2.3 Interviews With Project Researchers

To get a better understanding of the needs of the researchers, we conducted a series of semi-structured interviews with IoP researchers. The focus of the interviews was three-fold: We surveyed firstly, the kind of research data handled; secondly, the researchers' level of awareness and knowledge on RDM topics; and thirdly, their experience with RDM measures that were already put into practice, and perceived incentives and barriers regarding the implementation of further RDM methods. A question flow was developed in accordance with the three thematic clusters defined above. The

[5] NFDI4Ing, "Members and Participants", accessed January 13, 2023, https://nfdi4ing.de/about-us/members-participants/.

interview series took place from April to June 2021 and comprises 28 interviews with researchers from all cluster research domains. In the call for the interviews and the selection of interview participants, particular care was taken to ensure that the range of available research data and research cultures was covered as broadly as possible, which led to the subsequent insights.

2.4 Derived Requirements

In the following, we discuss the requirements that we derived from the interview series for creating and implementing a practicable DMP template for the IoP and projects of similar scope. We distinguish between content-related requirements (CR), i.e., implications regarding the elements of the DMP, and organizational requirements (OR), i.e., it is detailed which further aspects have to be considered when putting the work with DMPs into practice.

CR1: Elaborating on relevant topics. Templates provided by large organizations are often intended to be as generally applicable as possible. Subsequently, the included questions are kept generic, and the covered topics inevitably lack potential project- or discipline-specific focus. However, there may be specific RDM topics inherent to the project environment and area of research that are of particular relevance for making the research data understandable, and which should therefore be addressed in more detail in a customized DMP template. For some disciplines, there may already exist suitable templates.[6]

CR2: Ensuring comprehensibility and unambiguity. Existing templates often presume knowledge about vocabularies such as metadata, persistent identifiers, or the FAIR principles, which are not always common among researchers. Furthermore, we found that even basic terms such as 'research data' may leave room for different interpretations, and even more so in an interdisciplinary environment. A suitable DMP template should account for different levels of RDM proficiency in researchers as well as for various notions of important terms.

OR1: Top-down structure specifications. Without clear instructions from the management on responsibilities, hand-in schedules, and review processes, many researchers have indicated that they see neither the possibilities nor the reasons for creating DMPs. This applies in particular to projects where different institutes work together. The implementation of a DMP template should therefore be well coordinated and communicated by the project management.

OR2: Reducing overhead. A frequently expressed sentiment among the interviewed researchers was that DMPs were mainly perceived as an additional workload, which may have undesirable consequences on the completion behavior. Methods should be

[6] Science Europe. "Guidance Document Presenting a Framework for Discipline-specific Research Data Management." 2018. https://doi.org/10.5281/zenodo.4925907.

taken into consideration of how the amount of time needed can be reduced. Approaches to make the user experience more pleasant such as gamification may also be promising.[7]

OR3: Automation and integration with further services. For large-scale projects in particular, the need for reducing overhead wherever possible is also inherent to the work of data stewards and project managers. It is not feasible to manage the implementation, support the researchers, and evaluate DMPs without using suitable DMP tools and RDM infrastructure. This can, for example, include a platform for managing and evaluating all DMPs, functionalities to automatically fill in specific answers or to alert when fields are not filled in correctly, or export services. Additionally, by integrating working with the DMP with RDM services that substantially support researchers with their projects by assisting them with disseminating, reusing, and collaborating on research data, the usefulness of the DMP will be enhanced. This may include the linkage to repositories and archives, or suggestions of fitting standardized vocabularies.

3 Designing an IoP-specific DMP

As a foundation for the IoP-specific DMP, we used three existing DMPs. Two of them, the DMP provided by the Deutsche Forschungsgemeinschaft (DFG) which is used as a checklist for research proposals, and the RWTH DMP template, are intended to be as generally applicable as possible. Hence, we used both as input to focus primarily on the general structure of the IoP DMP and less on the exact phrasing of the questions (CR1). The third DMP we used as a foundation was the one specific to the Fuel Science Center (FSC) cluster of excellence at RWTH Aachen.[8] We used this DMP as an example for how to incorporate research specific topics and questions into the IoP DMP.

3.1 Structure of the DMP

With our findings we constructed the IoP DMP to orient towards four main sections:

General research project information. General information on a research project is available right from the beginning of project in most cases and it is also essential for locating the corresponding data within the IoP research cluster. The information contains, for instance, involved research departments or personnel responsible for the actual research data.

[7] Jüptner, Patrick, Manuela Dalibor, Ute Trautwein-Bruns, und Bernhard Rumpe. "Schulungskonzept zu Git und GitLab–Gamification zum besseren Lernen". In *E-Science-Tage 2021. Hrsg. Heuveline, Vincent, und Nina Bisheh,* 21–35. Heidelberg: heiBOOKS, 2022. https://doi.org/10.11588/heibooks.979.c13715.

[8] Hausen, Daniela, Jessica Rosenberg, Ute Trautwein-Bruns, und Annett Schwarz. "Data Stewards an der RWTH Aachen University – Aufbau eines flexiblen Netzwerks". *Bausteine Forschungsdatenmanagement*, no. 2 (2020):20-28. https://doi.org/10.17192/bfdm.2020.2.8278.

Detailed description of the research data. This part of the DMP focuses on the information about the actual contribution and research data of a project. This is done by asking researchers to provide tags describing the research data of a project,[9] but also provide details for defined data-sets including the contained types of research data, their formats, and involved software tools. This section of the DMP is usually the most active since research data is created continuously throughout the lifetime of a research project.

Storage, documentation, and other characteristics. Knowledge on storage locations and documentation of research data becomes especially important further down a project's life cycle. This is the case when older data is reused within the project to contribute to new research data. Additionally, decision on storage technologies or location oftentimes are not defined at the beginning of a project and might change frequently. Analogously, we found that the documentation of research data often is postponed to later stages of a project. Finally, characteristics of the data include standards or norms for data which are often not part of the usual documentation.

Legal obligations, knowledge exchange, and afterlife. The final part of the DMP includes information which in a lot of projects is determined in later stages or even at the end of a project. Legal obligations include, for instance, hints to personal data or any data related Non-Disclosure Agreements (NDAs). Especially the latter often restrict the exchange of research data as well as re-usability. Knowledge exchange includes questions on possible publications, possible usage in other research domains, and any existing time-dependent restrictions enforced by an NDA. In addition, the questions refer to archiving the data, its locations, as well as the applied technology.

3.2 Question Flow of the DMP

In order to refine the structure of the DMP and to improve the flow of questions, the IoP template makes use of conditional questions. We found that defining conditions for the order of questions, i.e., depending on previously given answers to a question following questions might be omitted purposely or raised differently, reduces the time to fill out the DMP. Moreover, it offers the opportunity to create several paths from the start to finish depending on the given answers.

The conditional flow of questions has several goals. Not only does it speed up the process of answering the DMP questions, it also reduces possible confusion caused by misleading questions which are seemingly out of context. Additionally, it provides a way to further customize the IoP DMP in such a way that it is possible to raise more

[9] Vairavasundaram, Subramaniyaswamy, Vijayakumar Varadharajan, Indragandhi Vairavasundaram, and Logesh Ravi. "Data mining-based tag recommendation system: an overview". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5, no. 3 (2015): 87–112. https://doi.org/10.1002/widm.1149.

detailed questions for certain topics without necessarily increasing the overall number of questions.

Raising more detailed questions depending on the answers given, leads to the final characteristic of the IoP DMP template. The DMP should not only serve as a way to document research data, in addition it should guide researchers through the process of RDM. To do this, questions are formulated in such a way that they guide researchers towards a certain direction, e.g., check if research data on subject already exists or provide answers which suggest to the researchers that they have to deal with certain challenges including the definition of metadata for their data or think about publications and the afterlife of their research data.

4 Implementing the DMP Template in the IoP

For the implementation of RDM within the IoP we concentrated on the requirements of subsection 2.4.

We developed a DMP template for IoP researchers which dedicates questions specifically to documenting and handling industrial data, e.g., machine data, or business-sensitive data from industry partners (CR1). For construction of the DMP we specifically focused on including help texts and examples for all terms that the interviews revealed to be unclear among researchers (CR2). The priority was to counteract any potential misunderstandings of these terms.

Additionally, steps towards the planned roll-out are closely coordinated with the cluster management (OR1). The goal is to avoid any unnecessary confusion which might be caused by the need of re-introducing structures or guidelines on how to include the DMP into regular workflows within the IoP.

Finally, we are exploring two methods for reducing the time required to complete the DMP. First, by adding a set of predefined response options to questions with free text answer fields, e.g., by surveying the most frequently handled data types and making them available as selectable response options. Second, by introducing a condition-based structuring at suitable places within the questionnaire, i.e., the appearance of questions is based on previous answers (OR2).

The following sections detail the different aspects by introducing the opensource tool Research Data Management Organiser (RDMO), explaining the process of creating the content of the questionnaire, and outlining the plan on how to establish RDM in the IoP with the help of RDMO.

4.1 Introducing RDMO

RDMO is a tool that supports to systematically plan, organize, and implement RDM throughout the life cycle of a research project.[10] For the IoP RDMO is utilized (1) to organize the internal structure of RDM with respect to different research projects and how they are mapped to RDMO, (2) to manage the IoP-specific DMP, (3) and to provide an interface between researchers and the DMP.

RDMO provides various management functionality to organize RDM. In Figure 1 this internal management structure is depicted. Projects can be organized hierarchically, i.e., a project can be assigned to another project as parent. As presented in the figure, each project has a set of members, which is also effected by the hierarchy. Members can take roles within a project. Available member roles are *Owner*, *Manager*, *Author*, and *Guest*. These affect the permissions of each project member, e.g., a guest is only allowed to view the answers of a DMP but cannot answer any DMP questions. Moreover, memberships and roles in RDMO persist through the project hierarchy, that is, a member of a parent project is automatically a member of any sub-projects while retaining the assigned role in the parent project. Finally, a project has an assigned DMP as shown in Figure 1.

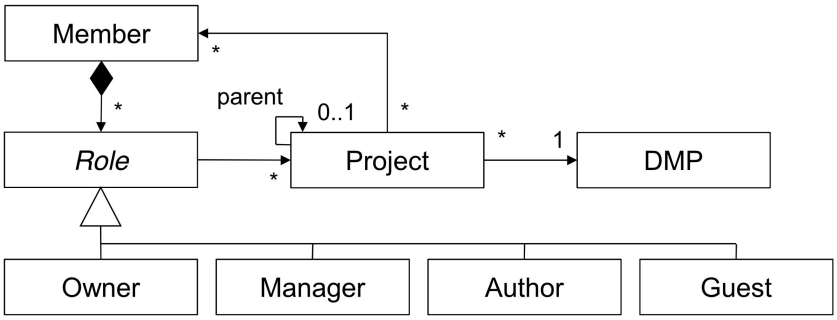

Figure 1: The class diagram presents the simplified structure of how RDMO manages projects and members. Each project has a set of current members. Members of projects have different roles within each project. The assigned role impacts the project-wide permissions of a member. Furthermore, projects can be structured hierarchically. Finally, a DMP is assigned to each project.

Figure 2 details how RDMO handles DMPs by presenting a simplified structure of how they are managed within the application. Each DMP consists of a set of questions which need to be defined. There are two types of questions, general questions and data-set specific questions. *General questions* refer to a research project in general, e.g., who the responsible persons for a project are or where and how a project will be archived. These questions only need to be answered once per research project. On the other hand, *data-set specific questions* refer to specific sets of research data within

[10] RDMO, "Research Data Management Organiser", accessed December 7, 2022, https://rdmorganiser.github.io/.

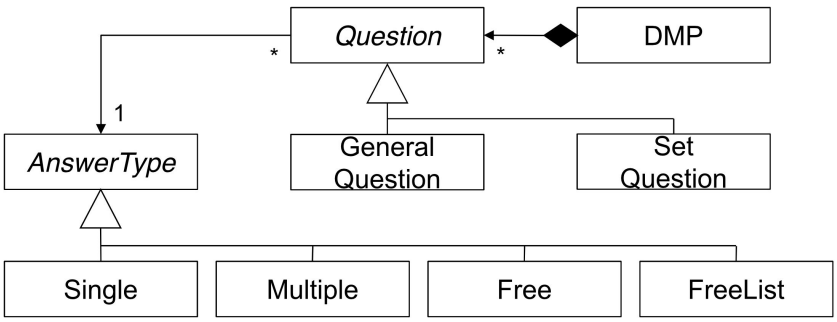

Figure 2: A class diagram modeling the simplified structure of a DMP in RDMO. Each DMP consists of a set of questions which either refer to a research project in general or to specific research data sets. Each question has a specific answer type. Possible types are single and multiple choice, free text, and list of free text.

a project. Such a set can, for instance, consist of a software implementation or lab results. The interface of RDMO gives researcher the opportunity to define these sets and afterwards answer the respective questions. Data-set specific questions must be answered for each set that is defined.

In RDMO questions can have different types of answers. Each question can either be answered in a single or multiple choice manner or by free text. Single and multiple choice answers can also include free text choices. Additionally, there exists the possibility to offer a list of free text answers, e.g., by providing the information on responsible persons as a list of single free text answers. Furthermore, RDMO allows to create conditions for the flow of questions. Consequently, it supports the concept of conditional questions as described in subsection 3.2.

Finally, RDMO offers a *Snapshot* service which allows to take snapshots of the current state of a DMP within a project. Snapshots contain all answers at the time they were taken. Each has a name and a description. In RDMO snapshots can be used to persist a project-specific history. Furthermore, snapshots can be used to set the current state of a DMP to the state stored in the snapshot providing a way to easily undo recent changes.

4.2 Establishing RDMO within the IoP

The IoP is internally structured into several research divisions called workstreams. Each workstream can be divided into additional internal research projects. Researchers of the IoP can participate in multiple of these projects.

To map the real world structuring of the IoP into RDMO we are planning to utilize the project hierarchy, membership, and roles functionalities provided by RDMO (cf. Figure 1). A simplified concept of this mapping is depicted in Figure 3.

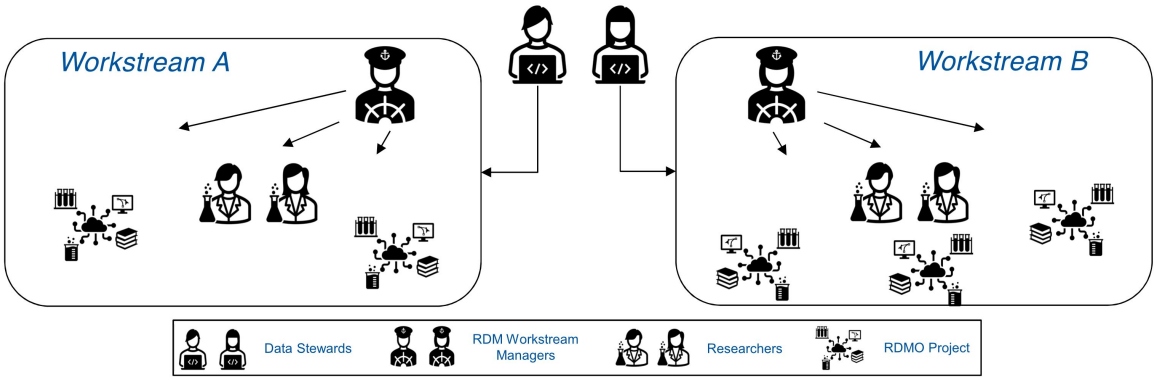

Figure 3: The simplified concept of mapping the IoP structure to RDMO. At the top the data stewards create projects in RDMO for each workstream. Members of these projects are the stewards and workstream RDM managers. The managers create projects which are sub-projects of the workstream project in RDMO for each of the workstream internal projects. The assigned researchers are members of respective projects in RDMO.

At the top of the hierarchy are the *data stewards*. They initially create the projects in RDMO which represent the IoP workstreams. Members of these projects are the data stewards and workstream RDM managers which are responsible for the general RDM in their respective workstream. Their role in the RDMO project is *Manager*. With the knowledge of the internal structure of their workstreams, the managers are able to create workstream-specific RDMO projects, mapping the real work structure to RDMO and adding the corresponding researchers as *Authors*. Each of these projects is a sub project of the respective workstream project in RDMO.

This structuring in RDMO reduces the overhead for the data stewards and avoids a single bottle neck when it comes to project creation and member management. Furthermore, it empowers the RDM managers of each workstream to easily manage their internal project structure and distribution of researchers and their roles in these projects.

In order to handle the handing-in process of DMPs the current concept makes use of the snapshot functionality of RDMO. In particular, the IoP management will schedule specific dates for handing in these snapshots, e.g., for the annual evaluation of workstreams. Nevertheless, it should also be possible for researchers to additionally specify their own schedule for snapshot creation to foster a flexible internal project-specific RDM process.

5 Conclusion and Outlook

A comprehensive introduction of RDMO and the DMP template was conducted in form of an IoP-wide workshop. A first iteration of researchers handing in DMPs is currently in progress.

While we generally recommend the proposed requirements for developing practicable DMPs in large-scale projects, it may not always be feasible to develop an own DMP template, depending on the resources of the project and the suitability of already existing, potentially discipline-specific, templates. On the other hand, a more granular staffing with data stewards in other projects may allow for an even higher level of customization and less top-down approaches in implementation. Further, while RDMO was the tool of choice for the IoP, as an instance was already embedded in the RDM infrastructure of the RWTH Aachen University, for other projects, a more comprehensive guide for an adequate choice may be required.

In further steps for the IoP, the submitted DMPs will be analyzed, and on that basis, the template will be iteratively refined regarding its design and its implementation. Additionally, the evaluation will serve as a resource to obtain a better picture of the state of RDM within the IoP, e.g., statistics about the types of research data, how research data is mainly disseminated, or the degree to which research data currently matches the FAIR principles. Further investigations of the realization of integrated services are planned, such as an automated pipeline from RDMO to other RDM tools such as to the RDM platform Coscine.[11] Finally, the information provided by the DMPs should not only be used for documentation, but also made available for researchers through services to find research data within the IoP in order to increase the usefulness for researchers and promote active engagement with RDM.

6 Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2023 Internet of Production – 390621612.

[11] RWTH Aachen University, "Coscine", accessed January 13, 2023, https://coscine.rwth-aachen.de/.