
Institutional research data management

Findings from the development and introduction of holistic research data management tools

Luca Leipold^{1,*}, Janine Straka^{2,*} and Kyanoush S.Yahosseini^{1,*}

¹Information and Research Data Management, Robert Koch Institute

²Department of Information Sciences, University of Applied Sciences Potsdam

*These authors contributed equally to this work

More and more research institutions commit themselves to follow data policies when managing their research data. With these policies, they aim to make their research data discoverable, accessible, interoperable and reusable. To support the implementation of these so-called FAIR principles, a suitable technical infrastructure is required. To provide such an infrastructure technical tools that enable easy metadata capture and facilitate research data management are needed. Although tools have already been developed to that end, there is still a lack of integrated platforms that cover the entire lifecycle of research data.

Here we present a prototype of such an integrated platform and describe our learning's from the implementation of this platform at the Robert Koch Institute.

In contrast to previously developed tools, our platform, the DataLinker, captures few data by itself, but mainly connects existing tools and services. More specifically, the DataLinker integrates metadata collected from adjacent systems such as the Research Data Management Organiser (RDMO). This integrated metadata is then linked to all research data associated with a project. This low-threshold approach allows for an easy management of research data at any stage of development. By providing a search interface and contact information, the DataLinker allows research data to be easily located, shared and reused.

We developed the platform to support researchers throughout the research data lifecycle. To do this, we worked closely with researchers during development, requesting and implementing their feedback. However, as we rolled out the platform, we found that requirements to such a platform go beyond the perspective of the researchers' data management needs. More specifically requirements from the perspective of an individual researcher are fundamentally different to the requirements when handling research data in an institutional context. For a platform to actually be used, it has to demonstrate clear added value besides its benefits to research data management. These benefits have to encompass not only the researchers but also the institutional service providers perspective. Providing benefits to the processes of these service providers, such as the IT department, legal department and the data protection office, are significantly relevant to the acceptance of such a platform.

Our work shows how project and research data processes are intertwined and cannot be considered independently. Therefore, platforms for simple research data management need to integrate and map both processes. We conclude that researchers need to be supported in both areas to enable them to follow institutional data policies.

1 Introduction

Proper research data management is becoming the focus of research institutions in the field of quantitative research around the world. As a result, more and more research institutions commit themselves to follow data policies when managing their research data. With these policies, they aim to make their research data discoverable, accessible, interoperable and reusable. To support the researchers in implementing these so-called FAIR principles [1], a suitable technical infrastructure is required. This infrastructure needs to include technical tools which match the researchers' everyday working life. These tools need to allow for easy research data management while providing a direct benefit, for example they should enable the researchers to easily capture, update and keep track of research and metadata. Although various tools have already been developed for this purpose, there is still a lack of integrated platforms that cover the entire lifecycle of research data.

Here we present a prototype of such an integrated platform and explain our experiences from the implementation of this platform at the Robert Koch Institute (RKI). The RKI is the national public health institute Germany [2], consisting of more than 1400 employees. The researchers work in very heterogeneous projects and tasks while handling very diverse research data. Historically (the RKI has been founded in 1891 [2]) additionally hamper the structured and open management of research data. As a result there is no platform or infrastructure to get an overview of all current and past projects and their associated research data so far. This is problematic as making research data findable is the first step towards a FAIR research data management.

Another challenge is the growth of data volume. Exponential growth of data volume brings conventional approaches to a limit, as possibilities for data storage, backup and long-term archiving must keep track of the increasing data column. For example, at the RKI approximately 80-100 terabytes of data are currently generated per year. Without proper research data management keeping track of this enormous quantities of data is next to impossible.

We aim to overcome these problems by introducing a new easy to use software platform for research data management which is developed according to the researchers' needs.

In this paper we first introduce our platform while describing the two components of the platform and their interactions in detail. Then we take a more theoretical perspective and describe the challenges we encountered and the lessons learned when introducing the platform to the institute. We suggest that these challenges can not be overcome by mainly technical means but that technical platforms and administrative processes and needs have to be considered side by side.

2 Technical development

We developed a novel integrated platform to guide manage researchers data through the whole research data lifecycle. Our platform consists of two components 1. the open source tool Research Data Management Organiser (RDMO) [4] and 2. a newly developed tool the DataLinker [9]. Both components fulfill different complementary tasks in our platform.

RDMO aims to simplify the process of managing and updating metadata. Hence in our platform RDMO is used to collect project metadata. Through a questionnaire based approach researchers use RDMO to gather various project related metadata, such as the title of their project and their collaborators. Using RDMO as a single-entry point and main hub for project related metadata avoids multiple entries of the same data when fulfilling the administrative demands of different departments in an institution.

The DataLinker automatically reads in data provided in RDMO. By showing a list of all projects which have been created in RDMO it provides a comprehensive overview of all research activity in an institution. Additionally imported projects from RDMO can be further enriched with additional data. Specifically, the DataLinker allows to connect projects with datasets and their physical location. This allows research data to be findable within an institution by searching for its metadata. Finally, the platform provides an easy import and export mechanism to public repositories. The DataLinker is not only limited to RDMO as a data source, but can read, use and digest information from other tools as well using public interfaces (Fig. 1). Hence our platform provides a standardised workflow from data management plan to data publication.

In the following section we introduce the two main components in more detail, while focusing on their technical implementation.

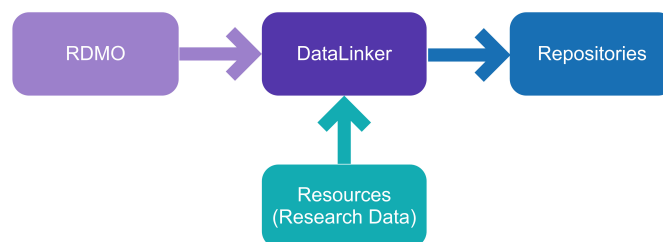


Figure 1: Components of our platform.

2.1 Research Data Management Organiser (RDMO)

The Research Data Management Organiser (RDMO) [4, 5] is a tool to create, update and work together on data management plans (DMPs) and to manage project related tasks. RDMO is an open source web application [6]. It is written in Python utilizing the Django package [7]. It supports a variety of different database systems, such as SQLite, MySQL, and PostgreSQL, as its main data storage. The public facing part of the tool relies on the AngularJS framework. Here we describe the general workflow of the tool followed by the changes we implemented to fit it into our platform.

The general workflow in RDMO focuses on the projects of a user and uses an interview style data entry format (a questionnaire) to collect data about the users' projects (see Fig. 2). A registered and logged in user can create a new project. For each newly created project a questionnaire has to be selected to indicate the type of the project. After selecting a questionnaire the user should answer questions in an interactive questionnaire. Some questions are dynamically generated based on the previous answers of the user. For example indicating that a project is funded by a third-party prompts questions which are related to this kind of projects. Questions which do not apply to a specific project are skipped. Also RDMO handles all answers as work in progress. That is, users can always skip questions or update their answers in a later stage of their progress.

RDMO allows to display questions and answers for each project in a structured way. These views can be used to automatically fill out pre-configured templates allowing for an easy way to inform third-party funders or handle administrative processes. To export answers into different file formats RDMO uses the Pandoc converter, a Python package [8], to export answers into different file formats.

The tool also provides an automatically generated list of tasks depending on given answers. These tasks indicate work items for the user. If such a task indicates that someone has to be informed about the state of a project, RDMO allows to generate and send a predefined e-mail with an attached view containing the requested information.

RDMO is a fundamentally collaborative tool. That is, a user can invite other researchers or collaborators to join a project and work on a questionnaire together. RDMO provides role and rights management for this purpose.

2.1 RDMO for institutional research data management

RDMO is extensible and highly configurable to the needs of an institution. To fit the needs of the Robert Koch Institute we adapted RDMO. We changed the look and feel of RDMO to match the look of other platforms of our institution. More importantly we adjusted questionnaires, implemented new tasks and views to mirror the processes and structures at the Robert Koch Institute. All modifications are published in GitHub [10].

We developed new questionnaires in RDMO to fit the needs of our platform. A generic questionnaire with up to 60 questions should be answered for every new project. The

MaMoDaR

Description

„MaMoDaR: Management Molekularer Daten im Research Data Life Cycle“ ist ein von der Deutschen Forschungsgemeinschaft (DFG) gefördertes Projekt mit einer Laufzeit von 24 Monaten (2019-2021). Das Projekt „MaMoDaR: Management Molekularer Daten im Research Data Life Cycle“ hat als Ziel, den effizienten und nachhaltigen Umgang mit Forschungsdaten zu optimieren. Es wird vom Robert Koch-Institut (RKI) in Kooperation mit der Fachhochschule Potsdam (FHP) realisiert. Hierzu dient die Entwicklung, Dokumentation und Veröffentlichung eines nachnutzbaren Konzepts sowie einer benutzerfreundlichen Softwarelösung. Die vom RKI entwickelte Software, der sogenannte DataLinker, unterstützt eine strukturierte Veröffentlichung wissenschaftlicher Daten nach den FAIR-Prinzipien.

Catalog

General questionnaire

Tasks

Task	Description	Time frame	Status
Contact IT department to acquire infrastructure resources	Please get in touch with your IT department to arrange that the required infrastructure resources are provided.		open
Inform Forschungskoordination (Fo, Research coordination)	Your planned project is a third-party funded project. Please provide a notice of third-party funding in accordance with the house order Fo. You can export the third-party funding report under the menu item Views.		open

Options

[Answer questions](#)
[Back to projects overview](#)

Export

[RDMO XML](#)
[CSV comma separated](#)
[CSV semicolon separated](#)

Import values from file

Figure 2: Example overview of a project in RDMO.

given answers in this questionnaire are the main source of metadata in our platform. We also developed three auxiliary questionnaires which are aimed at different project specific areas. These three questionnaires cover the handling of three areas: 1. data management plans, 2. administrative processes regarding the research project and 3. administrative processes regarding research data.

The first area comprises the core competence of the RDMO tool. Users are supported by the tool in creating data management plans allowing them to better manage their research data and to cover requirements for third-party funding applications. In the second area, users are supported in completing administrative applications that derive from their research project. These include cooperation agreements, third-party funding and data publication notifications. These applications can then be directly submitted to cross-sectional departments (such as the legal department, research coordination and the library). The third area includes processes that need to be initiated based on the nature of the research data. For example, in the case of research data which includes personal data, the data protection officer must be involved in coordination processes. Also when publishing sensitive data a dual-use analysis must be carried out. These three areas showcase one benefit of the platform: As many of the processes mentioned above require identical data. RDMO allows data provided in other contexts to be re-used to avoid reentering the same data multiple times.

Another major extension to RDMO was implemented for the so called tasks. Tasks in RDMO indicate work items which the user has to take care of. Tasks are generated

based on the user's answers in the questionnaires. For example, if a user answers the question "Does the dataset contain sensible data?", a task will be created which asks to contact the data security officer. In our platform we reused some of the tasks which were already implemented in RDMO. Additionally, new tasks were created to guide through internal administrative processes. For example based on the question "Is it a third-party funded project?" a task "Inform Forschungskoordination (Fo, Research coordination)" is generated. It includes a detailed description of the task such as "Your planned project is a third-party funded project. Please provide a notice of third-party funding in accordance with the house order. You can export the third-party funding notification under the menu item Views. On a regular basis, Fo should receive the notification at least 8 working days before the planned submission of the application to the funding agency!"

Our goal is to navigate the user of our platform through all mandatory administrative processes. This allows users to more easily comply with those procedures, as the system clearly shows them which applications have to be submitted and which formalities have to be observed. The answers given in the question allows to show only the relevant tasks with their corresponding time line. This makes it easier to meet deadlines, indicate which applications have to be filled and to whom an application should be made available.

2.2 DataLinker

The DataLinker is a web application that supports the capturing of project resources, the searching for project metadata, and the publishing of scientific research data in open repositories in accordance with the FAIR principles. The DataLinker has interfaces to adjacent systems and obtains project metadata primarily from RDMO via a REST API in JSON format. However it is not only limited to RDMO as a source of information. The DataLinker follows a layered architecture, it consists of three layers: a database-based persistence layer for data storage, a business layer implemented in the Java Spring Boot Framework [17] handling the application's logic and a web-based presentation layer implemented in Angular [18]. The business logic and fronted communicate by using a REST API [19]. In order to give the user the feeling of being in a single environment, the design of the DataLinker was based on the design of RDMO. The open source code for DataLinker is available at GitHub [9], a documentation [11] and a manual [13] have been published

The DataLinker provides three core functionalities to expand on the capabilities of RDMO. These three functionalities revolve around 1. search, 2. tracking of research data (so called resources) and 3. publication of research data. To provide these functionalities each project created in RDMO is mirrored into the DataLinker automatically, using the REST API of RDMO. In principle other providers of metadata besides RDMO can be integrated as well.

The DataLinker reconditions the collected metadata and provides a listing of all created projects with their relevant metadata. This metadata includes for example: contact person, title, project description and usage rights. The DataLinker allows resources, the

location and type of data sets, to be attached and tracked to a project. As a result users can easily get an overview of where data is stored, which licences are used and which type of data is been stored.













The three core functionalities of the DataLinker reflect in the tab-based user interface. The tab “Search” allows users to search for other projects and their associated meta- and research data via a keyword search (see Fig. 3). Search results can also be filtered via a faceted search by relevant metadata such as organizational unit, funder, or cooperation partner. This makes allows for different search strategies. For example, researchers interested in influenza can use the search to find already existing research data on their topic. They can also look up who has already worked on influenza and which data is available at which location. Data and also the expertise of others can be reused. Another example refers to department heads who can get an easy overview of all projects which have been carried out in their department, both historically and currently.

The screenshot shows the DataLinker interface for a project named 'MaMoDaR'. The top right corner indicates the contact person is 'leipoldi'. The main content area is divided into two parts: a detailed description of the project and a table of linked resources.

Project Description:

- Beschreibung:** „MaMoDaR: Management Molekularer Daten im Research Data Life Cycle“ ist ein von der Deutschen Forschungsgemeinschaft (DFG) geförderes Projekt mit einer Laufzeit von 24 Monaten (2019-2021). Das Projekt „MaMoDaR: Management Molekularer Daten im Research Data Life Cycle“ hat als Ziel, den effizienten und nachhaltigen Umgang mit Forschungsdaten zu optimieren. Es wird vom Robert Koch-Institut (RKI) in Kooperation mit der Fachhochschule Potsdam (FHP) realisiert. Hierzu dient die Entwicklung, Dokumentation und Veröffentlichung eines nachnutzbaren Konzepts sowie einer benutzerfreundlichen Softwarelösung. Die vom RKI entwickelte Software, der sogenannte DataLinker, unterstützt eine strukturierte Veröffentlichung wissenschaftlicher Daten nach den FAIR-Prinzipien.
- Kontaktperson:** Datendromedar
- Projektleiter*in:** Anna Gram
- Organisationseinheit:** MF 4
- Titel:** Management Molekularer Daten im Research Data Life Cycle
- Abkürzung:** MaMoDaR
- Schlüsselwort:** Metadaten, Forschungsdatenmanagement, Forschungsoutput, FAIR-Prinzipien
- Drittmittelprojekt:** ja
- Förderer*in:** Deutsche Forschungsgemeinschaft (DFG)
- Kooperationspartner*in (extern):** Fachhochschule Potsdam
- Speicherort:** S:\OE\MF4\Projekte\MaMoDaR
- Geplante Lizenz / Nutzungsbedingung:** Anders: Apache Lizenz 2.0

Projekt Datensätze:

Beschreibung	Quelle	Standort	Typ	Lizenz / Nutzungsbedingung	Aktion
Projektdatensatz MaMoDaR	Gruppenlaufwerk (S:\OE)	S:\OE\MF4\Projekte\MaMoDaR	Ordner	Nicht offen/Keine	 
Webaufruf des Projekts MaMoDaR	WebseiteURL	https://www.rki.de/mamodar	Homepage	Offen	 
Poster für die RDA Deutschland Tagung 2020	WebseiteURL	https://www.rda-deutschland.de/ecoster2020/ecoster_mamodar_rda-de-2020_2-1.pdf	Poster	Offen	 
Code	Git Repository	https://github.com/mamodar/	SoftwareCode	Namensnennung 3.0 (CC BY 3.0)	 
Code Dokumentation	WebseiteURL	http://datalinker.h2888668.stratoserver.net/doc/	Dokumentation	Namensnennung 3.0 (CC BY 3.0)	 
Manual	WebseiteURL	https://mamodar-docs-en.rki.de/docs/joinlatest/	Dokumentation	Namensnennung 3.0 (CC BY 3.0)	 

At the bottom right of the table, it shows 'Objekte pro Seite: 10' and '1 - 6 von 6'.

Figure 3: Example projects as shown in the DataLinker with its linked resources.

The tab “My Projects” allows users to attach resources (research data or data sets) to a specific project. This linkage again allows the connection between location of resources and the corresponding projects. The storage of the resources is not altered by the DataLinker, just a pointer to the location of the resources is saved and tracked. This allows the actual research data to remain on a local storage drive, being moved around or to be published to an external source.

The last tab “Publications“ supports the user in publishing data sets to external public repositories such as edoc [15] or Zenodo. Here edoc is the Robert Koch Institute’s own publication server which is based on DSpace [14] which is used in many different institutions. Zenodo is a generic repository which is financed by the European commission. Due to the modular structure of the DataLinker, the data export can also be easily implemented for other repositories. Before publication the user can review their already

collected metadata. After approving the publication metadata and research data is automatically transferred to the selected repository and published. If a DOI [12] is generated during the publication process it is automatically written back into the DataLinker.

3 Platform development

The development of our platform was guided by the needs of the researchers at our institution. To understand their everyday work and to assess their previous knowledge of research data management several approaches were used.

First we launched an institute-wide survey to gain insight into the extent to which researchers are already familiar with research data management and which tools they use for this purpose. We based our survey on the survey conducted at the Potsdam University of Applied Sciences and adapted the questions to our concerns [2]. The survey conducted in February 2020 (51 participants) showed that the majority (86%) of all survey participants (in departments who work with molecular data) deal with research data in their work context. However, experience in research data management is way less common (58% of all participants indicated that they have little to no experience). The survey also indicated that most researchers use digital tools to manage their research data which are not FAIR compliant. More specifically most participants indicated that they use familiar but ill-fitting tools such as the word-processor Microsoft Word and the spreadsheet software Microsoft Excel to handle their research data.

Second, we conducted expert interviews to better understand the users' everyday work and to adapt the tools to their usual work structures. We bundled our results from these interviews with the feedback of the surveys and created a list of requirements. The development of our platform followed this list. Afterwards, usability tests were carried out and adjustments were made to the system in close feedback loops with the researchers. The adaptations made to the platform were discussed in a workshop with internal and external participants. The resulting feedback then again was transferred to the development of the systems.

4 Institutional research data management

Platform requirements for research data management from the perspective of an individual researcher are fundamentally different to the requirements when handling research data in an institutional context. When introducing our platform to a wider audience we found that fulfilling all research data management requirements is not enough for a general acceptance of a system. For the tool to actually be used, it has to demonstrate clear added value besides its benefits to research data management. This is especially important within the framework of institutional research data management. Here it became apparent that the role of cross-sectional departments has to be given a large focus.

Such departments as the IT, the research coordination, the data protection officer and the legal department play a crucial role when introducing new platforms: Their support or rejection of a platform propagates towards the acceptance by the researchers. To generate such acceptance an advertised platform needs to provide a clear added value not only to the researchers but also to these service providers. Hence introducing any new research data management platform is a process which should be jointly initiated.

It is important to realise that research data management is only one module in an institute's landscape. Each department has its own established processes and vocabularies and a justification to use exactly these. Hence, another central finding is that research data management must be integrated into the established administrative processes to achieve wide acceptance. Especially project and research data management processes are closely intertwined, so they need to be considered together. To use RDMO as the single point of entry into research data management proved to be impracticable for our approach. In contrast we suggest to focus on linking the existing processes and their used (technical) tools. In this way, all concerns can be contextualised and applied in the respective expert systems. The necessary data should then be automatically transferred between these expert systems.

5 Conclusions

We developed a research data management platform which consists of two components 1. RDMO as a tool to capture and handle project related metadata and 2. DataLinker as a tool to make this metadata FAIR. We closely cooperated with interested researchers to develop and adapt our platform to their needs. Our experience during the introduction of the platform shows, that focusing on the users and the technical implementation of such a platform is not enough. Instead the introduction of a research data management platform also requires focus on administrative processes. That is why the structures of an institution needs to be addressed in detail and all people involved need to be identified and in-cooperated. To manage this enormous effort we suggest to not replace existing expert systems but to augment them with proper research data management platforms.

In summary, we identified two pillars that have to be given equal attention and are a prerequisite for the success of a project.

The first pillar is the needs of the researchers. Each platform has to be developed according to the users every day work. The goal aim here is to support the user of the platform while avoiding as much additional work as possible. Furthermore, a clear added value must be noticeable to the user, it has to become clear why it is worth to manage research data properly and what benefits the user can have by making research data available to others.

The second pillar comprises the organisational structure. New processes which are inevitably established by the introduction of new technical platforms should be based on existing processes. In particular it is important to include the cross departments in all

developments. These departments, such as the research coordination, the IT department, the data protection officer, the legal department and ethics committee, support scientists in their projects and are hence an integral part of the success of such a platform. Moreover close integration of research data management with project management is essential.

Acknowledgements

Our platform was developed as part of the project "Management Molekularer Daten im Research Data Lifecycle (MaMoDaR)". The project has been funded by the Deutsche Forschungsgemeinschaft (DFG), Germany - Project number 416783714. The Robert Koch-Institute and the University of Applied Sciences Potsdam participated in the realization of the project.



~~The authors thank Linus Grabenhenrich for his valuable advise during the paper creation.~~

Bibliography

- [1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. "The FAIR Guiding Principles for scientific data management and stewardship" *Sci Data* 3, 160018 (2016): 1-9. <https://doi.org/10.1038/sdata.2016.18>.
- [2] Website: https://www.rki.de/EN/Content/Institute/Mission_Statement/Mission_Statement_node.html (last visited: 09.03.2021)
- [3] Website: www.rki.de/mamodar (last visited: 09.03.2021)
- [4] Website: <https://rdmorganiser.github.io> (last visited: 09.03.2021)
- [5] Website: <https://rdmo.readthedocs.io/en/latest> (last visited: 09.03.2021)
- [6] Website: <https://github.com/rdmorganiser> (last visited: 09.03.2021)
- [7] Website: <https://www.djangoproject.com> (last visited: 09.03.2021)
- [8] Website: <https://pandoc.org> (last visited: 09.03.2021)
- [9] Website: <https://github.com/mamodar/datalinker> (last visited: 09.03.2021)
- [10] Website: <https://github.com/mamodar/rdmo-rki-catalog> (last visited: 09.03.2021)
- [11] Website: <https://github.com/mamodar/datalinker> (last visited: 11.03.2021)
- [12] Chandrakar, R. "Digital object identifier system: an overview." *The Electronic Library* (2006).
- [13] Website: <https://mamodar-docs-en.readthedocs.io/en/latest/> (last visited: 06.04.2021)

- [14] Website: <https://duraspace.org/dspace/> (last visited: 31.03.2021)
- [15] Website: <https://edoc.rki.de/> (last visited: 06.04.2021)
- [16] Arndt, O., Glatz, L., Hummel, B. et al. "Umfrage zum Forschungsdatenmanagement an der FH Potsdam : Projektbericht" Zenodo (2018) <https://doi.org/10.5281/zenodo.1161792>
- [17] Website: <https://spring.io/projects/spring-boot> (last visited: 06.04.2021)
- [18] Website: <https://angular.io> (last visited: 06.04.2021)
- [19] Website: <https://restfulapi.net> (last visited: 06.04.2021)