Research Data Management Organiser

A tool to support the planning, implementation and organisation of research data management

Claudia Engelhardt
University of Applied Sciences Potsdam
Faculty of Information Sciences
Kiepenheuerallee 5
14469 Potsdam
Germany
claudia.engelhardt@fh-potsdam.de

Harry Enke
Leibniz Institute for Astrophysics
Potsdam (AIP)
An der Sternwarte 16
14482 Potsdam
Germany
henke@aip.de

Jochen Klar
Leibniz Institute for Astrophysics
Potsdam (AIP)
An der Sternwarte 16
14482 Potsdam
Germany
jklar@aip.de

Jens Ludwig
Staatsbibliothek zu Berlin
Preußischer Kulturbesitz
Potsdamer Straße 33
10102 Berlin
Germany
jens.ludwig@sbb.spk-berlin.de

Heike Neuroth
University of Applied Sciences Potsdam
Faculty of Information Sciences
Kiepenheuerallee 5
14469 Potsdam
Germany
neuroth@fh-potsdam.de

ABSTRACT

The project Research Data Management Organiser (RDMO)[1] develops a tool to support the planning, implementation and organisation of research data management. The multilingual open source tool can be installed locally and adapted to institutional or discipline-specific needs with regards to contents. It provides interfaces to institutional authentication procedures. Key features of the first version, released in April 2017, include:

- the ability to continuously update and augment the information in the course of a project
- access for different stakeholders, such as researchers, project coordinators, the IT department, data managers etc., with customized views
- export formats for different purposes including data management plans (DMPs) according to funder requirements.

Planned future developments include, among other things, features to support actual data management, e.g. tasks (with deadlines and a reminder functionality) that can be linked with dedicated stakeholder responsibilities.

KEYWORDS

research data, data management planning, active data management, tool, organiser

1 INTRODUCTION

The basis for a successful long-term management and provision of digital research data is a thorough research data management throughout the whole project lifetime. The goal is to get FAIR data – data that are findable, accessible, interoperable and re-usable [6]. Data management is not solely a responsibility of researchers, but also of research institutions that have to provide the necessary technical infrastructure, consulting and support. The project Research Data Management Organiser (RDMO) developed RDMO as a tool that supports both researchers and institutions in the planning, implementation and organisation of research data management.

RDMO was funded by the German Research Foundation DFG and, in its first phase, ran from November 2015 to April 2017.

The paper outlines the starting points, summarises the conceptual and development work done so far, describes the tool's features and fields of application and gives an outlook on future developments.

2 DATA MANAGEMENT PLANNING – THEORY AND PRACTICE

2.1 Data Management Plans

The foundation for FAIR data is laid very early in the research process, namely in the planning stage. For this reason, data management plans have become a crucial element of data management policies of funding agencies, and subsequently for universities and other research institutions. DMPs vary in extent and detail; typical elements include statements about what data will be created with which methods, applicable policies, plans for sharing and preservation, data curation measures and respective

[1] http://rdmorganiser.github.io/en/

responsibilities, ownership and access conditions, restrictions (e.g. for legal or ethical reasons) and required resources [2, 3].

In recent years, tools to support the creation of DMPs according to funder requirements have been developed, the most well-known and popular of which are DMPonline[2] by the Digital Curation Centre (DCC) and DMPTool[3] by the California Digital Library (CDL).

2.2 Active Data Management

But creating a plan is just the first step – since a plan does not help much if the outlined data management measures are not put into practice. Data management is an ongoing effort which includes adjusting the DMP and the associated activities if necessary. To emphasize this continuous character, the term active data management has been introduced.

The growing awareness of the need for such an active data management shows, for example, in the Research Data Alliance (RDA) Interest Group on "Active Data Management Plans"[4], the requirement for EU Horizon 2020 projects to update their data management plans in case of significant changes [5] as well as in the joint activities of DMPonline and DMPTool to "reposition DMPs as living documents" [11] and "integrat[e] them into the broader ecosystem of data management infrastructure" [10].

The same motivation inspired the RDMO project.

2.3 Why another tool?

Both DMPonline and DMPTool, at least currently, have a strong focus on the requirements of research funders. Their main purpose is to discover, edit and fill in DMP templates of funding organisations, mainly from the UK (in the case of DMPonline) and the US (in the case of DMPTool). In most instances, these DMPs will be forgotten after the submission. Although active data management is considered in their future plans, it is not supported by the tools at the moment.

Both tools are centralised web applications. This is associated with the transfer of potentially sensitive information off-site. Also, it offers institutions only limited possibilities for customisation. Furthermore, DMPonline and DMPTool are not re-deployable without investing considerable effort.

All of these points are addressed in the RDMO concept.

3 CONCEPTUAL DESIGN

The design process of RDMO was guided by the following aims. The tool should:

- support the data management throughout the whole project lifetime
- enable users to gather and organise all information necessary for a sustainable data management
- involve all relevant stakeholders[5]
- be locally installable and configurable.

As crucial features required to realise these aims we identified:

- collaborative editing
- specific roles, rights and views for different stakeholders
- input via a structured interview; skipping of redundant or unnecessary questions based on given information
- output of gathered data in various forms, amongst these textual data management plans for different funder requirements
- possibility to adapt the contents (questions as well as answering options)
- tasks and reminders
- easy application and administration in different contexts (e.g. university, research institute, joint research project).

The conceptual design was developed mainly on the basis of three activities: 1) the further development of previous work to design a basic, generic questionnaire, 2) desk research, user tests and interviews on discipline-specific requirements and content, and 3) user stories to model the requirements and the associated functions for different stakeholders.

During the whole process, we were supported by a number of projects and institutions[6] who tested the tool at different stages and gave us valuable feedback and input.

3.1 Basic questionnaire

The major source for the development of the basic questionnaire was the *WissGrid-Leitfaden zum Forschungsdatenmanagement* [9]. This is a research data management manual and checklist produced by the project WissGrid (2009-2012).[7] In Germany, it has become one of the standard works on research data management.[8] In addition, we also looked at DMPs and similar material of other tools, institutions or funding agencies as well as checklists that one of the project members had developed for the use in consultations with researchers while working as a data manager at a Max Planck Institute. The latter already included tasks associated with certain questions to stress the active element in data management.

The basic questionnaire of RDMO is designed to cover all relevant data management aspects. In particular, it comprises of the following areas:

- general information about the project (including data management requirements from policies, e.g. of funders or the home institution)
- content classification
- technical classification
- data usage
- data storage and security
- collaborative work
- quality assurance

[2] https://dmponline.dcc.ac.uk/

[3] https://dmptool.org/

[4] https://www.rd-alliance.org/groups/active-data-management-plans.html

[5] See section 3.3 for a description of stakeholder roles and requirement.

[6] See http://rdmorganiser.github.io/en/cooperations/

[7] http://www.wissgrid.de/

[8] The WissGrid checklist again was inspired by previous work, too. In the introduction, the editors particularly refer to the DCC's Checklist for a Data Management Plan, v3.0 (http://dcc.ac.uk/webfm_send/431).

- costs
- metadata and referencing
- legal and ethical issues
- long-term preservation (incl. selection and appraisal).

Users can either use this pre-assembled generic questionnaire – or can build their own. The latter allows the customisation of the tool in terms of discipline, method, institutional context and / or other relevant factors.

3.2 Discipline-specific requirements

To find out more about the discipline-specific requirements of such a tool, the project investigated two disciplines as an example: the social sciences and astrophysics. In a first step, we undertook desk research on data management practices and requirements in these two fields [the main sources being 4, 7, 8]. The results were then verified in expert interviews with a data manager of TwinLife[9], a 12-year longitudinal twin study on the development of social inequality, and the working group on optical solar physics[10] at Leibniz Institute for Astrophysics Potsdam (AIP). The interviews took place in May and June 2016 and included testing a draft version of the generic RDMO questionnaire. Finally, the results were also discussed in a breakout session of an RDMO workshop held in Potsdam in July 2016, where our previous findings were verified.

The results suggest that – on the mid-range or coarse-grained level of granularity which most DMPs or tools, including our own, cover – the need for the customisation of questions according to specifics of certain disciplines is given only for single questions or topics. In astrophysics, for example, usually no personal data or data protected by copyright is gathered or produced, so that these topic areas can be omitted. For large, quantitative studies in the social sciences, it is common to outsource the data collection and parts of the data preparation to external survey institutes. Therefore, a DMP tailored to the social sciences should address the topic of external survey institutes and the steps in the research process these are responsible for. But as mentioned before, in general this applies only to a few selected areas of the DMP. A discipline-specific customisation of larger parts or the whole DMP would only be useful on a very detailed, fine-grained level. Then, however, the questions become so specific to a narrowly defined set of use cases that this does not seem to be a sensible strategy for the vast majority of user groups.

With respect to the answering options as well as help texts and links for further information provided by a DMP or RDM tool, however, our interview partners as well as the workshop participants saw great potential benefit in customised options, e.g. by preparing sets of options that refer to data types, methods, tools, standards, vocabularies etc. commonly used in the discipline in question or suggesting repositories for certain disciplines or data types where possible. This would make it more comfortable for researchers to answer the questions. In addition it would help to standardise the answers which then would make it easier for an institution or IT department to collect information across a number of projects to, for example, assess what infrastructure resources (e.g. storage) need to be provided. However, the design of questionnaires tailored to different disciplines or sub-disciplines is a task that cannot be accomplished by a project like RDMO. It can best be tackled by the scholarly communities in question, and RDMO offers the means to implement this.

3.3 Stakeholders and user stories

To find out more about the requirements of the different data management stakeholders, we collected user stories. Instead of just simply postulating requirements, user stories reflect the perspectives of different actors (in my role as ...), describe what activity they want to use the tool for (I want to ...) and then indicate the purpose (to get the benefit of...) to put the activity into context [1].

In total, we compiled about 70 user stories for the following stakeholder roles:

- author (most common)
- infrastructure provider (second most common)
- superior (third most common)
- data manager
- guest
- manager
- IT administrator
- developer
- IT support
- reviewer

We cannot go into detail at this point, but shortly present the categories that were aggregated from the user stories, each with one or two examples.

Collaboration

- As author, I want to invite other persons to my DMP as reader or author, so that they can contribute.
- As superior, I want to be able to read and approve DMPs, to fulfil my controlling duties.

Usability, input assistance and templates

- As author, I want to use templates and recommendations from my institution and funding organisation, to know what to focus on to fulfil their requirements.
- As author, I want to have predefined selections of useful, correct and standardised answering options where possible, to save time because then I do not have to think of my own ones first.

Versioning

- As author, I want to access older versions of my DMP, to undo incorrect entries.

[9] http://www.twin-life.de/en

[10] http://www.aip.de/en/research/research-area-cmf/cosmic-magnetic-fields/solar-physics/optical-solar-physics?set_language=en

Adaptability of questions and answering options

- As infrastructure provider, I want to be able to define selections of useful, correct and standardised answering options, in order to be able to aggregate user inputs for easier analysis.

Logic of questions

- As author, I do not want to answer questions that can be identified as irrelevant on the basis of my previous entries, so that I can concentrate on the questions that are important for my project.
- As infrastructure provider, I want to offer questions and answers in different levels of granularity, so that plans can be made according to the needs, abilities and most useful level of detail for a certain project.

Task administration

- As infrastructure provider, I want to connect certain questions and answers to data management tasks for different actors and roles, so that all stakeholders are aware of their tasks and responsibilities in the process.
- As infrastructure provider, I want to be able to view the answers for relevant questions, in order to plan accordingly (e.g. how much storage space needs to be provided).

Review

- As reviewer, I want to have access to the information in RDMO, to evaluate if the planning and implementation of a project's data management are / were carried out properly.

APIs and export functions

- As author, I wish that information I or others entered in this or other systems can be migrated, so that I do not have to enter everything from scratch again.
- As author, I want to be able to export my answers in a machine readable form and link them to other systems, to be able to use the information entered in other RDMO installations (at other institutions).

4 TECHNICAL DEVELOPMENT AND FUNCTIONS[11]

RDMO is implemented as a web application based on the Python framework Django[12] and the JavaScript framework AngularJS[13]. It is an open source tool licensed under Apache Version 2.0[14] and is available on GitHub[15].

The information in RDMO is organised along projects. It is up to the users to define what project means in their specific context. In most cases, a project in RDMO will represent a 'real life' research project, but it might also relate to a subproject, one particular survey or study or a range of studies inside a project, or others.

The RDMO implementation is multilingual, currently available languages are German and English. It is designed to support internationalisation, so that more languages can easily be implemented in the future.

In the project life span of 18 months we achieved the goals set in the project proposal and even realised some additional features. However, not all requirements and desirable features that were identified by the user stories and feedback from testers could be implemented by the end of the project in April 2017.

In the following, we give a brief description of the features of the first version, released at the end of April 2017.

4.1 Input

Information can be entered in a web interface and edited by different stakeholders. It is collected via a structured interview that guides users through all relevant topic areas. Based on the answers given, redundant or irrelevant questions are skipped. Depending on the type of question/answer, different widgets are used (e.g. radio buttons, check boxes, drop-down lists, rulers, free text). Controlled vocabularies or predefined sets of answers are used when available. However, as the basic questionnaire is generic, this is extensible in questionnaires tailored to particular disciplines or fields of study. Snapshots can be made at any time to freeze and document the state of the information about a project at a given point in time.

4.2 Output

There are several kinds of output. On the project level, the information previously entered can be aggregated into textual views (e.g. DMPs to be used for project proposals). These views can be defined as required. There is a to-do list of data management tasks to be performed and a reminder function that notifies the responsible parties of upcoming assignments.

On a departmental or institutional level, information can be aggregated across projects, which is useful for a number of purposes, e.g. to derive the demand regarding infrastructure resources or to get an overview of the types and amount of data produced at the institution.

4.3 Setup and operation

One aim was to design RDMO in a way that makes it easy to install and to customise in various contexts (e.g. universities, research institutes, departments of universities or research institutes, libraries, larger joint research projects, research groups etc.). Accordingly, the tool is flexible in several respects. In order to allow institutional branding, the user interface can be freely customised. The content can also be tailored completely to suit context-specific needs. This applies to the questions and answering options as well as to help texts, templates for the output of DMPs and tasks.

Another important point is the ability to integrate it into the local infrastructure. RDMO can be linked to the institutional

[11] For the whole chapter see also http://rdmorganiser.github.io/en/software/
[12] https://www.djangoproject.com/
[13] https://angularjs.org/
[14] https://www.apache.org/licenses/LICENSE-2.0
[15] https://github.com/rdmorganiser/rdmo

authentication and authorisation infrastructure, namely an LDAP system or a Shibboleth federation.

5 OUTLOOK

In addition to the demo version provided by the RDMO project (hosted by Leibniz Institute for Astrophysics Potsdam (AIP) and available at https://rdmo.aip.de), several institutions have set up local instances, among them Karlsruhe Institute of Technology, University of Stuttgart, University of Konstanz, University of Duisburg-Essen, and the Göttingen eResearch Alliance. Several more have expressed interest in doing so.

Further development of RDMO is planned in the future, the second project phase will start in autumn 2017. Some of the aforementioned institutions will closely collaborate with the project in this respect.

The range of functionalities will be extended, e.g. by the possibility to upload relevant documents, such as codebooks, metadata documentation or applicable guidelines, and a commenting functionality. Existing functions will be refined. The focus of these projected activities lies on features supporting the implementation and organisation of the actual data management throughout the project lifetime, e.g. roles, tasks, modules for cost estimation and ingest-process. The interoperability both between different RDMO instances and with external services such as re3data[16] or research information systems will be enhanced. Set-up, upkeep and integration in different institutional environments will be made easier with standardised installation, integrated maintenance mechanisms and extended support of authentication and authorization procedures. Exchange and cooperation with different stakeholder groups will be continued and intensified with the objective to build an active user community dedicated to the continuous joint distributed development as the basis for a sustainable future of RDMO. This includes tutorials and workshops for different user groups as well as the extension of our cooperation with existing data management and infrastructure initiatives

[16] http://www.re3data.org/

REFERENCES

[1] Scott W. Ambler 2014. User Stories. An agile introduction. In: *Agile Modeling*. [online] Available at http://www.agilemodeling.com/artifacts/userStory.htm [Accessed 16 March 2017]

[2] *ANDS Guide. Data Management Plans.* 11 January 2017. [online] Available at: http://www.ands.org.au/__data/assets/pdf_file/0007/731779/Data-management-plans-.pdf [Accessed 16 March 2017].

[3] *DCC. Data Management Plans.* [online] Available at: http://www.dcc.ac.uk/resources/data-management-plans [Accessed 16 March 2017].

[4] Harry Enke & Johannes Wambsganß 2012. Astronomie und Astrophysik. In: Neuroth et al. (eds.): *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme.* Boizenburg: Verlag Werner Hülsbusch, pp. 275-293. http://nestor.sub.uni-goettingen.de/bestandsaufnahme/kapitel/nestor_bestandsaufnahme_014.pdf [Accessed 16 March 2017].

[5] European Commission, Directorate-General for Research and Innovation 2016. *Guidelines on FAIR Data Management in Horizon 2020.* Version 3,0. July 2016, p. 5. [online] Available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf [Accessed 16 March 2017].

[6] FORCE11. *The FAIR Data Principles.* [online] Available at: https://www.force11.org/group/fairgroup/fairprinciples [Accessed 16 March 2017].

[7] Michael Häder 2009: *Der Datenschutz in den Sozialwissenschaften.* (RatSWD Working Paper No. 90). http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_90.pdf [Accessed 16 March 21017].

[8] Uwe Jensen 2012. *Leitlinien zum Management von Forschungsdaten. Sozialwissenschaftliche Umfragedaten.* (GESIS Technical Reports 2012|07). http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-07.pdf [Accessed 16 March 21017].

[9] Jens Ludwig & Harry Enke (eds.) 2013: *Leitfaden zum Forschungsdaten-Management.* Glückstadt: Verlag Werner Hülsbusch. jens.ludwig@http://www.wissgrid.de/publikationen/Leitfaden_Data-Management-WissGrid.pdf [Accessed 16 March 21017].

[10] Stephanie Simms et al. 2016a. The Future of Data Management Planning: Tools, Policies and Players. In: *International Journal of Digital Curation*, Vol. 11, Iss. I, p. 208-217, here: p. 1. DOI: http://dx.doi.org/10.2218/ijdc.v11i1.413

[11] Stephanie Simms et al. 2016b. Roadmap: A Research Data Management Advisory Platform. In *Research Ideas and outcomes*, Vol. 2: e8649, p. 2 DOI: http://dx.doi.org/10.3897/rio.2.e8649