

D5.5 Actionable Modular Data Management for brokering distributed resources

Author(s)	Adil Hasan (Sigma2), Bert Meerman (GGF), Hanne Moa (Sigma2), Ulf Jakobsson (GU/SND), Joakim Philipson (SU)
Status	Final version
Version	1.0
Date	11/Nov/21

Document identifier:	
Deliverable lead	Sigma2
Related work package	WP5
Author(s)	Adil Hasan (Sigma2), Bert Meerman (GFF), Hanne Moa (Sigma2), Ulf Jakobsson (GU/SND), Joakim Philipson (SU)
Contributor(s)	
Due date	31/Aug/21 (delayed to 15/Nov/21)
Actual submission date	
Reviewed by	Hamish Struthers (SNIC), Mari Kleemola (TAU)
Approved by	Hamish Struthers (SNIC)
Dissemination level	PU
Website	https://www.eosc-nordic.eu/
Call	H2020-INFRAEOSC-2018-3
Project Number	857652
Start date of Project	1/09/2019
Duration	36 months
License	Creative Commons CC-By 4.0
Keywords	

Abstract:

This deliverable describes work done on implementing in four different Data Management Plan tools support for machine-actionable data management plans (maDMP) that is necessary for actionable modular data management for brokering distributed resources. The work covers the applications of the maDMP and lessons learned.



Copyright notice: This work is licensed under the Creative Commons CC-BY 4.0 licence. To view a copy of this licence, visit <https://creativecommons.org/licenses/by/4.0>.

Disclaimer: The content of the document herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the document is believed to be accurate, the author(s) or any other participant in the EOSC-Nordic Consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the EOSC-Nordic Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the EOSC-Nordic Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

Table of Abbreviations	3
1. Executive Summary	5
2. Introduction	5
3. Data Management Plans	6
3.1 History	6
3.2 Stakeholders	7
4. Tools for DMPs	9
4.1 DMPonline	9
4.2 Data Stewardship Wizard	11
4.3 EasyDMP	12
4.4 SND Checklist	13
5. Machine-actionable DMPs	14
5.1 Applications of the RDCS	15
5.1.1 SU-EOSC Nordic 5.3.2 maDMP project	15
Challenges with maDMP's	17
Potential benefits compared to other approaches	18
Recent and future development	18
5.1.2 Machine actionable plans in EasyDMP	19
Supporting RDCS	19
Interoperability with the DSW and NIRD services	20
5.1.3 Machine actionable SND checklist	22
6. Discussion	22
6.1 The DMP Lifecycle	22
6.2 Lessons Learned	23
6.2 The wider context	24
7. Next steps	25
8. References	26

Table of Abbreviations

Table 1: Abbreviations used in this document.

Abbreviation	Explanation
API	Application Programming Interface
CDL	California Digital Library
DCC	Digital Curation Centre
DMP	Data Management Plan
DOI	Digital Object Identifier
DSW	Data Stewardship Wizard
ELIXIR-NO	Norwegian node of the European life science infrastructure for biological information
FAIR	Findable Accessible Interoperable Re-usable
H2020	Horizon 2020
ICPSR	Inter-university Consortium for Political and Social Research
JSON	JavaScript Object Notation
KTH	Royal Institute of Technology, Stockholm
maDMP	machine-actionable Data Management Plan
NIRD	Norwegian Infrastructure for Research Data
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
ORCID	Open Researcher and Contributor ID
PID	Persistent Identifier
RDA	Research Data Alliance
RDCS	RDA DMP Common Standard
RDM	Research Data Management

SE	Science Europe
SEO	Search Engine Optimization
SND	Swedish National Data Service
SNIC	Swedish National Infrastructure for Computing
SU	Stockholm University
SUNET	Swedish University Computer Network
VR	Vetenskapsrådet (<i>Swedish Research Council</i>)
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations

I. Executive Summary

In studying “Actionable Modular Data Management for brokering distributed resources”, we have focused on the Data Management Plans as a tool to manage the relevant data components for a particular research.

Our findings have allowed us to make the following conclusions::

- A Data Management Plan (DMP) is a crucial tool to define and describe the data part of a research project. It contains a wealth of information that is of value to many stakeholders.
- DMPs that are text driven (essay DMPs) are traditionally used by funders to judge the feasibility of the research, but extracting information from these DMPs is not a simple task.
- A machine actionable DMP (maDMP) is strongly preferred over a traditional DMP, as an maDMP adheres to the FAIR Principles and can be processed, evaluated, shared and reused in an effective way by the different stakeholders.
- Different DMP tools and templates have been independently developed, due to differing stakeholder groups and intended audiences. As this use of DMPs has increased, standardization and interoperability of tools/templates has become more complex and has been recognized as important for reuse.

2. Introduction

Practically every research project uses and produces data. It forms the heart of the project and effective handling, organizing, structuring and storing of the data over its life is essential to ensure the successful completion of the project. These activities, collectively called data management, need to be considered

beforehand in order to efficiently manage the data. The goal of the DMP is to provide a set of data management procedures agreed by the project owners that, if followed, will enable the data to remain useful throughout its lifetime. Currently, most DMPs are written as essay form text documents (for example <https://zenodo.org/record/5500391>).

The DMP contains a lot of information that can be useful to administrators (for example to understand which communities plan to use which resources), researchers (for example, who wish to create a similar project and would like to know the storage profile in order to estimate their own profile), storage providers (for example, who wish to understand how much storage has been requested by projects in the coming year) and many more stakeholders. However, extracting information from the current documents meant for human consumption is time-consuming and error prone, and with the need for each project to provide a DMP manual assessment will not scale.

The potential of the DMP and the value in it being machine actionable has been recognised by [6] and the Research Data Alliance (RDA) working group on DMP Common Standards created a schema (called the RDA DMP Common Schema, RDSC) that enables machine actionable DMPs to be created.

In this deliverable we report on the activities we have undertaken to implement the RDSC in the DMP tools we are currently supporting. We also describe the challenges we faced and future work on these activities. Our work has resulted in a publication [4] and several presentations at the RDA maDMP workshop and other workshops.

3. Data Management Plans

This section starts with a short description of the history of the DMP, followed by a subsection describing the stakeholders that can take advantage of a well designed DMP.

3.1 History

The first “DMPs” were used in the late 1960s in complex engineering and aeronautical projects, outlining expected research and development activities.¹ In the 1970s and 1980s, DMPs were used as active project management tools to deal with “data management requirements during data collection and/or analysis stages.”² These early DMPs were created to solve complexities identified in data creation/collection, processing and short-time data storage. The use of “DMPs” at that time was impelled by researchers who knew the requirements of their own projects.

The development of DMPs during the 21st century arose from public policies in two main areas: economic policies as well as eResearch. In 2004, the Organisation for Economic Co-operation and Development (OECD) published a declaration³ that “...recognised the beneficial impact of open access data...”, resulting in the formation of a working group that pointed out that OECD countries suffered low returns on public funding due to a near-absence of data reuse. They pointed out that the responsibility to share data should be placed on the researchers. They did not directly advocate the use of DMPs, however “...various aspects of data access and management should be established in relevant documents, such as [...] grant applications [...]”⁴. One of the more important reports from the eResearch section discussed the almost

¹ Small, N et.al. (2020). *A Review of the History, Advocacy and Efficacy of Data Management Plans*. International Journal of Digital Curation. <https://doi.org/10.2218/ijdc.v15i1.525>

² *ibid.*

³ Committee for Scientific and Technological Policy. (2004). *OECD declaration on Access to Research Data from Public Funding*. Paris, France:organization for Economic Cooperation and Development.

⁴ Pilat, D., & Fukasu, Y. (2007). OECD principles and guidelines for access to research data from public funding. *Data Science Journal*, 6, OD4-OD11

complete absence of institutional and government funding for data repositories. This report⁵, commissioned by the Joint Information Systems Committee, UK (JISC), was very concerned that the storage media that were used at that time, had life spans of only a few years and thus the data would soon have disappeared forever if nothing was done. The report suggested that strategies for data management, sharing and preservation should be developed. However, the report suggested that institutions and government, not the individual researcher, should be responsible and that identifying and archiving data of potential future value was of essence.

In 2005 the National Science Board (USA) recommended that the funding body, the National Science Foundation, require a peer-reviewed DMP with all grant applications.⁶ Simultaneously, a report⁷ in the UK recommended that funding bodies should require DMPs with funding applications. Depending on funding bodies, the content of the DMPs varied from a full data lifecycle approach, specific requirements due to the funder's own objectives, to a focus on data sharing ignoring the other aspects of the data life cycle. None of the recommendations were backed by any type of quantified evidence nor were any reference made to the research communities.

From 2010 onwards voices were raised to align research practice but also the content of DMPs. In the US most DMPs covered topics like data archiving and sharing, but not so often topics like ethics, data capture, and intellectual properties; the same thing happened in the UK. Also more and more funding bodies required DMPs with the applications which was a catalyst for university libraries to create and promote DMP tools and processes.⁸

In both the UK (the Digital Curation Centre, DCC) and USA, DMP tools were developed to assist researchers fulfil the requirements of DMPs in their funding applications. In the UK DMPonline was developed, while DMPTool followed shortly after in the USA.

Turning our attention to the Nordics, and taking Sweden as an example, work on information materials regarding DMPs started in 2010 at the Swedish National Data Service (SND). Initially this was in-house documents describing what a DMP was, what it could contain and why it was (supposedly) good to have one. These documents were in line with what other Research Data Archives in Europe and USA, as well as organisations like the Digital Preservation Coalition, Inter-university Consortium for Political and Social Research (ICPSR) and DCC, stated. Over the following years, less focus was placed on funder requirements and more on what was useful for the researchers, the administration as well as SND itself (since SND were the organisation who in the end could end up with curating and disseminating the data). One of the first documents on DMPs was unofficially made available in 2011 after a call from the Swedish Research Council in 2012, where they asked for a Data Publication Plan. SND published information on data management and publication on the web, however it took until 2013 before the first DMP-document was officially published on SNDs web. Later on that year, the first public presentation by SND staff on an international conference was made (19th EAA Annual Meeting, Pilsen, the Czech Republic, *Data Management Planning, What it is and how to do it*⁹). New versions were created over the following years where more information was added including explanatory texts on why different topics were important as well as for who. New areas were added due to ethical and legal problems that might arise for researchers depending on type of data

⁵ Lord, P., & Macdonald, A. (2003). *E-Science curation report: Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision*. Digital Archiving Consultancy Limited.

⁶ National Science Board. (2005). *Long-lived digital data collections: enabling research and education in the 21st century* (Technical Report). Arlington, Virginia: National Science Foundation

⁷ Digital Archiving Consultancy, Bioinformatics Research Centre, & National e-Science Centre. (2005). *Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models*.

⁸ Bishoff, C., & Johnston, L. (2015). Approaches to data sharing: An analysis of NSF data management plans from a large research university. *Journal of Librarianship and Scholarly Communication*, 3(2).

⁹ <http://legacy.ariadne-infrastructure.eu/resources-2/presentations/ariadne-workshop-eaa-2013/>

and project. It was in 2016 that the first “Checklist for Data Management Plan”¹⁰ was officially published on SNDs website, and it has been further updated with new versions after that, both to make sure that the checklist was in line with requirements from Horizon 2020 and the Swedish Research Council, but also in the later years with the Core requirements of Science Europe.

3.2 Stakeholders

Each DMP contains information of value to more than one stakeholder. There are several ways of grouping stakeholders into larger categories. For this report, we have identified the following five core groups of stakeholders and what they might be interested in from a DMP. The identification of these groups is based on SNDs work on material about DMPs as well as (in early 2020) several meetings and workshops with researchers, RDM administrators and other administrative staff at universities where the participants discussed the different groups and what information from a DMP they might be interested in/have use of (a more complete list of stakeholders can be found in, for example [4]):

1. The *Funder* who wishes to ensure the data produced by the funded projects has the most extensive use possible, ensuring that the most value for money is obtained from the funds provided. The funder's needs are expressed mainly in their DMP guidelines that are used to provide DMP templates. Nordic Funders often delegate the assessment of DMPs to the research institution (see stakeholder 3 below). Reasons for this are, for example (i) lack of expertise within the specific area, (ii) not enough staff to assess the DMPs and (iii) some funders are of the opinion that it is the universities responsibility to have a DMP for each project to make sure that the data is handled correctly and made available at the end of the project. More and more funders are requiring that the data produced is to be published according to the FAIR principles¹¹ and with Open Access (As open as possible, as closed as necessary). In the case where the DMP is reviewed, the Funder relies on the assessment of the project peer reviewers.

2. The *Researchers* (and other *Data Producers*). In this group we have several roles: the primary researcher who is responsible for the project but also the project leader, data managers, the research group, research institutes and finally the secondary user (who intends to reuse the data). These different roles wish to make sure that the data management elements have been identified as much in advance as possible so that the project does not stall for lack of resources, access or data loss. For this stakeholder group, the DMP is an important organizational document, defining who in the group has which area of responsibility. This is particularly important if there are several researchers in the project or if the project is a collaboration with other parties, but also an advantage if more researchers come in later. If a DMP exists, and all project members commit and adhere to the DMP, it becomes easier to have control over the project outputs. One must keep in mind that research teams have a completely different need to keep track of a DMP. Compared to the individual researcher, teams need to coordinate and have control over the handling of data. With a DMP one can also plan where the data should be made available and then ensure that PIDs are correctly assigned so that the citation can be done. Also secondary users who wish to find, evaluate or reuse data for new projects can use the information in the DMP (for example, researchers in the Norwegian Meteorological Institute have expressed an interest providing Digital Object Identifiers, DOIs, for DMPs so they can be shared). They might also have questions about its content (sensitive data, ethical issues in reuse) and how they should handle such data themselves in the future.

3. The *Institution/University*. This is one of the larger groups with roles/entities like the archive, IT-department, librarians, administration, legal officers, data controller, data protection officers, Grants and

¹⁰ <https://snd.gu.se/en/manage-data/guides/dmp-checklist>

¹¹ <https://www.go-fair.org/fair-principles/>

Innovation Office. Here we also have the *RDM administrators* (such as curators, data stewards and reviewers) who wish to ensure the data produced by researchers at their institution are of sufficiently high quality that they can be reused and may also use the DMP as an information collection tool. Since the university is responsible for the research conducted at the university, a DMP can be used to promote quality research data and metadata but also that the handling of the data is carried out correctly (legal concerns, archiving etc.). However, that requires, at least in part, that the researcher/research group work together with the RDM administrators. From the university's point of view, there are several other reasons for the creation of a DMP: for example (i) to learn what data is created and improve findability/searchability, (ii) to make sure that the data is handled correctly and thereby fulfilling archiving requirements (law and legislation) (iii) enabling data reuse, (iv) being able to allocate resources for storage, (v) managing access and security around sensitive data. For these purposes they also need DMPs to be easy to review and evaluate, preferably, to the extent possible, by automated processing, in order to allow for an expected upscaling and increase in the number of DMPs.

4. The *Infrastructure*. Here we have entities like Research Data Repositories, National Archives, Computing facilities like SUNET (Swedish University Computer Network) and SNIC (The Swedish National Infrastructure for Computing, that makes available large-scale high-performance computing resources, storage capacities, for Swedish research), and Data Centres that manage data. These organisations are currently working to enable data reuse and they can use information from DMP to plan storage needs, cost calculations, development of services, for analysis and management of security aspects etc.

5. The *Society*. In this group we have the taxpayer (government), authorities, corporations, trade and industry, journals, but also the social benefit that a project can provide. It is in everyone's interest that tax-funded research is as resource-efficient as possible. If researchers can re-use data there will be an increase in efficiency - for example a reduced cost of data collection. Open access to research data is therefore important. Also, it is important to think about the searchability of DMPs when creating them (i.e. it is important that the data management plans are also FAIR!) Society, as the 'highest authority' has the possibility to change or create new guidelines as well as update policies. Many research projects run for a long time, so research should be documented such that information is understandable. Transparency is important as society should gain access not only to how research data is handled but also to have the opportunity to request the data. What research is done can be made visible with a DMP, which enables the use of the data for other things than research, e.g. community planning (infrastructure, etc).

4. Tools for DMPs

Differing stakeholder needs have resulted in a large number of tools existing for creating data management plans (see [1] for an extensive review of the existing tools). Researchers are free to choose the DMP tool that best fits their needs. In this section we describe the four tools that we are currently using. All the tools described below can also be installed and run locally.

4.1 DMPonline

DMPonline¹² is used in the work described in Section 5.1.1. It is a tool (see Figure 1) for supporting DMP creation and management offered by the Digital Curation Centre (DCC), based in the UK but now serving

¹² <https://dmponline.dcc.ac.uk/>

individual users and institutions globally.

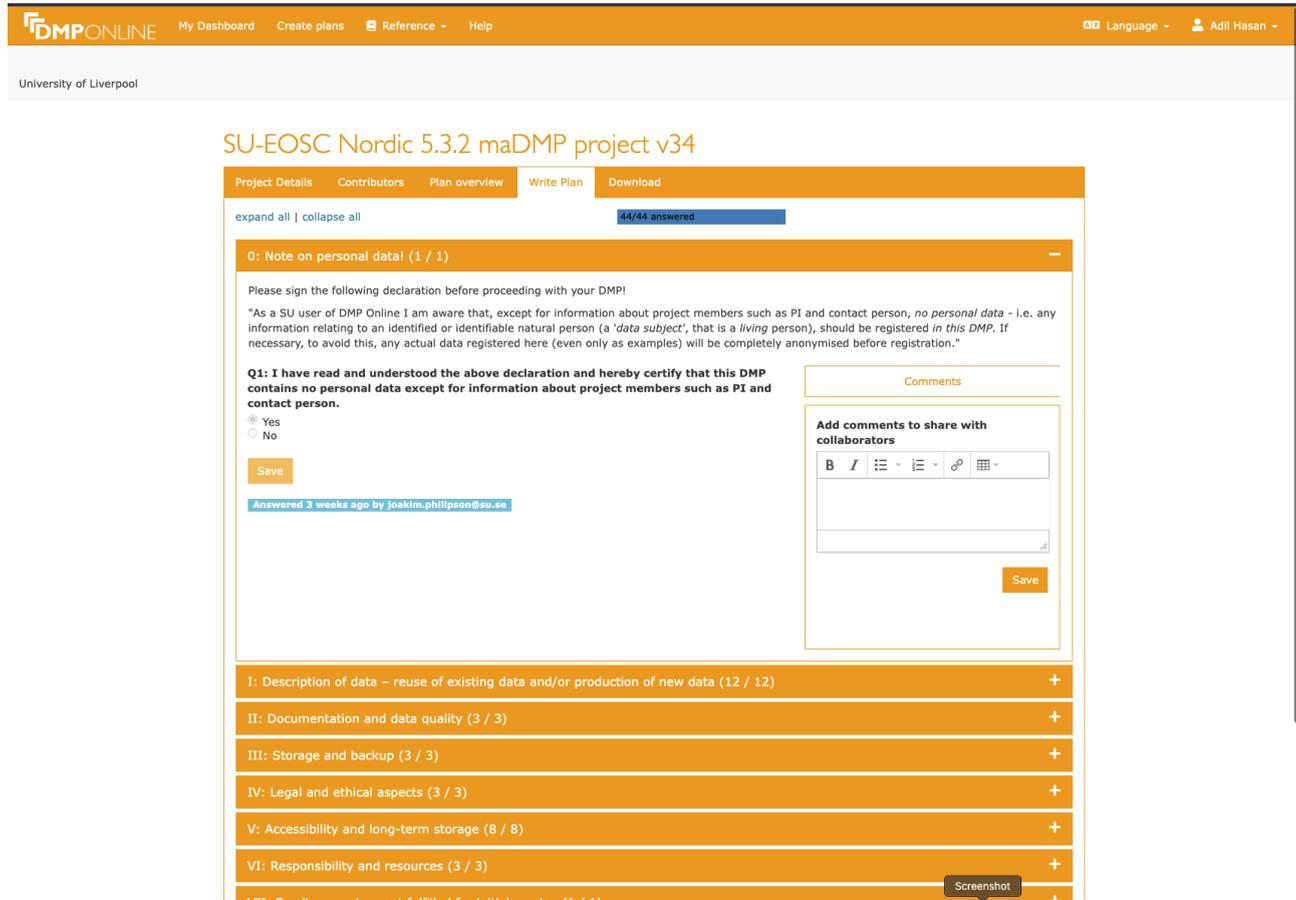


Figure 1. Screenshot of DMPonline showing the SU-EOSC Nordic 5.3.2 template described later.

The DCC and the California Digital Library (CDL) are collaborating on a development plan, DMPRoadmap, that is being successively implemented in their tools (DMPonline and DMPTool) respectively at a different pace.

DMPonline offers access to over twenty public funder templates, that the individual user can choose between for the creation of their DMP. These different templates tend to vary considerably in what information they ask for. What they have in common is that they are almost exclusively free text based, requiring shorter or longer essay answers to questions. To make these machine-actionable, if at all possible, would at least involve recourse to natural language processing (NLP) techniques.

In addition to the public funder templates, DMPonline offers the possibility of creating your own local institutional templates, which however are only fully accessible to your own local users, affiliated at your institutions and then with some further functional limitations compared to the funder templates.

Both partners of DMPRoadmap also offer APIs for extracting filled DMPs in JSON format. For more details on APIs see Challenges with maDMPs in Section 5.1.1.

4.2 Data Stewardship Wizard

Figure 2. Screenshot of the DSW showing the subsets of the knowledge model that can be selected.

The Data Stewardship Wizard¹³ (DSW, see Figure 2) is used in the work described in Section 5.1.2. It was initially developed for the European Infrastructure for Life Sciences (Elixir) project and is heavily used by the life sciences community. It continues to strongly support DMPs for the life sciences, but also supports DMPs from a much wider range of disciplines. The DSW calls DMP templates knowledge models and DMP template designers (such as RDM administrators) create knowledge models for their domain that satisfies the relevant data management guidelines supplied by the institution and/or funding agency. The DSW allows a hierarchy of knowledge models to be created allowing knowledge models to inherit from other models. Knowledge models that meet the Horizon 2020, Science Europe and Elixir data management guidelines exist as well as many others. A researcher can then select a knowledge model and the parts of the model they need in order to build a compliant DMP. For example, in Figure 2 the researcher has selected the Science Europe and Swedish Research Council subsets and will only be presented with questions that conform to those data management guidelines.

¹³ <https://ds-wizard.org/>

The DSW aligned with the FAIR principles early on and the core knowledge models provide questions that address each of the FAIR principles. The DSW makes it easier to create structured knowledge models that can lead to more machine-actionable and consumable plans.

4.3 EasyDMP

The screenshot shows the 'Choose a template' page in the EasyDMP application. The page header includes the 'easy.DMP' logo and navigation links for 'Your plans', 'Help', 'Admin', and 'Log out'. The main content area displays a table of available templates. Below the table, there are navigation controls for the list, including 'Previous', '1', '2', and 'Next'. At the bottom of the page, there are logos for 'easy.DMP', 'EUDAT', 'UNINETT', and 'sigma2', along with links for 'User Guide', 'About', 'Support', 'Terms of use', and 'Privacy policy'.

Template	Version	Description	Published	Retired
Horizon 2020	1	Simplified template based on Horizon 2020 guidelines.	2018-01-25 14:04:07Z	No
Horizon 2020 Expert	1	A shorter template based on Horizon 2020 that assumes knowledge of data management.	2019-04-10 07:43:09Z	No
Linear Example 2017	1	Template to demonstrate a linear-flow data management plan	No	No
Minimal branching template	1	A simple branching template, to test branching issues.	No	No
NINA data management plan	1	Template for the Norwegian Institute for Nature Research	2021-02-22 13:21:02Z	No
Norwegian Institute of Bioeconomy Research	1	Template for the Norwegian Institute for Bioeconomy Research.	No	No
RDA DMP CS test	1	This template will be used to improve support for the interchange format RDA DMP Common Standard.	No	No
Science Europe	2	Template for data management plans based on the Science Europe guidelines.	2019-04-08 07:10:35Z	No
Sigma2 Data Management Plan	1	Data management plan for Sigma2 resource allocation.	2020-09-09 05:02:12Z	No
UIO PSI: Enkel datah�ndteringsplan for forskning som ikke er Helseforskning	1	Template for PSI	No	No

Figure 3. Screenshot of EasyDMP showing the list of available templates.

EasyDMP¹⁴ (see Figure 3) is used in the work described in Section 5.1.2. It is developed and managed by Uninett Sigma2 in Norway. An instance accessible to Norwegian and international researchers is also run on the Sigma2 infrastructure. One of the original goals of EasyDMP was to provide a simpler interface for creating DMPs with questions containing canned responses, or with controlled vocabularies. Another goal was to integrate with the services provided by the Norwegian Infrastructure for Research Data that is managed by Sigma2, such that a researcher would fill in the DMP, answering the appropriate questions for services. Once the plan is approved this will trigger the allocation of those services. In common with the other tools, a RDM administrator can create a DMP template for a community. New templates can be derived from existing templates through cloning. RDM administrators have the flexibility to create highly

¹⁴ <https://easydmp.sigma2.no/>

structured DMPs that are more easy to transform into machine-actionable plans, or highly narrative plans that would be easier for a human to comprehend.

EasyDMP currently provides templates for the Science Europe guidelines, Horizon 2020 and the Sigma2 guidelines for new projects.

4.4 SND Checklist

The SND checklist is used in the work described in Section 5.1.3. It is an extensive document, which can be downloaded in both Swedish and English (<https://snd.gu.se/en/manage-data/guides/dmp-checklist>).

Formats are pdf and for an editable version as docx or rtf.

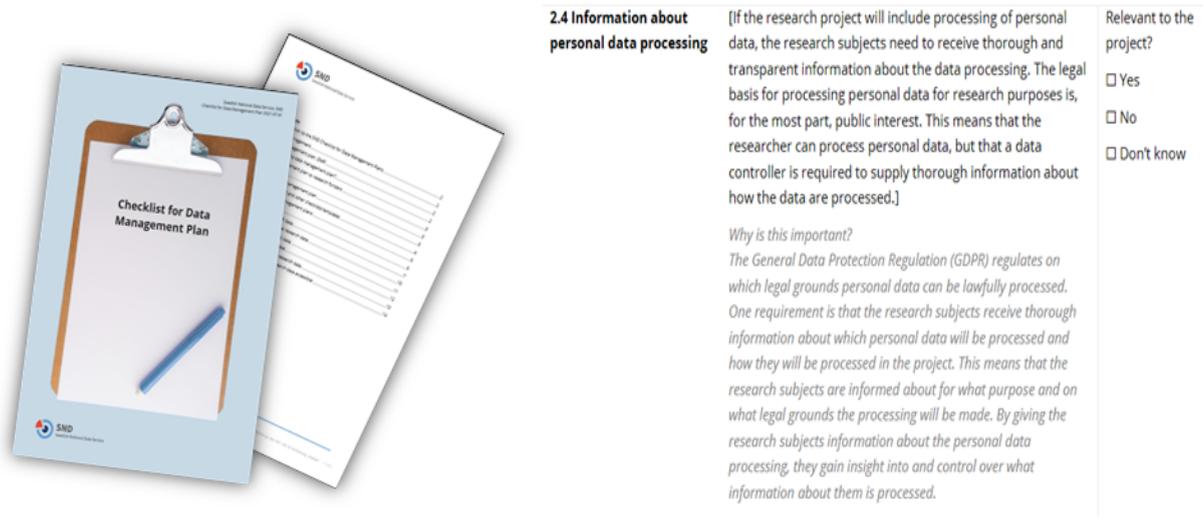


Figure 4. The SND checklist. The screenshot on the right is of a question in the checklist.

The checklist is updated regularly and complies with, among others, the recommendations from Horizon 2020 and Science Europe, which the Association of Swedish Higher Education Institutions (SUHF) and the Swedish Research Council also follow. The SND checklist may differ from the recommendations in, for instance, design and terminology, but contains the same information. Another major difference is that the SND checklist has, to a much greater degree, separated the questions and has more explanatory text (see Figure 4).

The checklist was created as a response to both the realisation of its usefulness and later, to a funding call from the Swedish Research Council where they asked for a Data Publication Plan and pointed the researchers to SND. Apart from several questions to be answered there are also sections explaining what a DMP is, why one should create one, the difference between the checklist and funder requirements, the relation between DMPs and the FAIR principles, etc. However there are thoughts of making it machine actionable (see section 5.1.3).

The intended audience is the researchers themselves, giving them the opportunity to discover what they ought to do to manage their data well. The idea is that the researcher either by himself/herself or with the support from SND/the (Swedish) universities Data Access Units goes through the checklist, noting which parts are applicable for the project. The researcher then starts writing the DMP and subsequently updates it over the project period.

5. Machine-actionable DMPs

Until recently a DMP was considered to be a document containing details on how the data were to be managed. In many cases the researcher used a DMP tool to select the appropriate template for the funder (or agency requiring a DMP) and filled in the template's series of free-text questions covering the full data life-cycle from the creation of the dataset through to use, reuse and deletion of the dataset. The resulting document would then be submitted to the requestor who would review and approve, deny or request modifications to the plan. The requestors often required the DMP to be updated mid-way through the project.

Over time we have noticed some issues:

- For a given template, for example when assessing DMPs for NIRD storage requests, the quality of the DMPs varied considerably which made assessment of the plans a time-consuming process. Questions could be answered with terse one-sentence responses, or with text that did not answer the question. In some cases the quality could be improved by providing stronger guidance in the help, providing examples, or making the questions more specific.
- Reuse of the information in the DMP is very difficult. The free-text nature of the document makes it difficult to identify and extract information that could be used by another service. For example, a storage service would need to work hard to consume the response “about six terabytes now and 15TB for the rest of the project” to the question “How much storage space do you need per year?”.
- The actual data management diverged from what was described in the DMP. The fact that the DMP existed as a document disconnected from the data management processes meant that it was very easy for the intentions described in the DMP to no longer correspond to reality unless a researcher put effort into updating the DMP.

These issues impact the usefulness of the DMP for the identified stakeholders. A more structured DMP that could be consumed by a machine (a maDMP) would help to address these issues and would allow the stakeholders to access the information relevant to them. Miksa in [4] proposed a set of principles for maDMPs, one of which was the use of a common data model. An RDA working group on DMP Common Standards¹⁵ was setup to develop a common data model, (the RDCS), and produce a JSON reference implementation¹⁶ of the model.

It is worth mentioning that Kim [2] also puts forward the potential use of DMPs as a *discovery tool*, but here also the needs may differ between different stakeholders. For example, for RDM-administrators / curators / data stewards, DMPs might serve to help discover in which data repositories the datasets described by DMPs are to be found, allowing them to make predictions about the ensuing data quality based partly on what is already known about the level of curation, metadata standards, data policies of those repositories. For researchers themselves, on the other hand, searching data repositories and publication databases seem more likely to be the primary means of discovery.

¹⁵ <https://www.rd-alliance.org/groups/dmp-common-standards-wg>

¹⁶ <https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard>

Even social media such as Twitter may be more important in this respect than DMPs, which are often not made publicly available. However, if public DMPs can bring together datasets and research articles / publications, as [2] also seems to suggest, then perhaps they could also serve as a more general discovery tool for different stakeholders. In part, this might be achieved by allowing DMPs to also hold a list of references / publications that is continuously updated, maybe even registering relation types (as for related identifiers in Zenodo) describing how different references are related to the research project described in the DMP.

On the other hand, it is important to avoid forcing researchers to fill out the same information and metadata repeatedly in different documents / web forms. To the extent possible, a maDMP should be pre-populated by pulling in information from other relevant sources, or by allowing for inference from a minimal set of (meta)data entries in the DMP. For example, common metadata standards to be used in a research project can be inferred from a statement of which data repository/ies that will be used for making data accessible.

Regarding *namespaces* / vocabularies / ontologies to be used in DMPs, in particular for the purpose of serving as a *discovery* tool, use of *schema.org* which was developed as a SEO tool deserves to be explored, as suggested by [2]. Another vocabulary that might serve the same purpose is perhaps *FaBio* (<https://sparontologies.github.io/fabio/current/fabio.html>). Apart from potentially making DMPs more discoverable, using common namespaces and metadata standards also *in DMP templates* (which RDCS does not yet require) would likely contribute to making DMPs both more interoperable and machine-actionable.

5.1 Applications of the RDCS

In this section we describe how the RDCS reference implementation was used to address the needs of the RDM Administrators, the NIRD Archive and the Producers and Consumers of datasets through the integration of the RDCS in the DMP tools we are using.

5.1.1 SU-EOSC Nordic 5.3.2 maDMP project

In this project Stockholm University (SU) developed a custom DMP template in DMPOnline that was named the *SU-VR template*, as it is based on the Science Europe and Swedish Research Council (VR) sections and questions, but the output of which would be machine-actionable to the extent that it would validate against the RDCS maDMP-schema. To achieve this proved to require also some further processing of the output via API. However, the purpose of the project has been threefold, where the first of these was the most important to us:

1. To ease the *administrative burden* on our SU researchers to fill in a DMP, by means of preset multiple-choice answer options, drop-down menus, radio buttons etc., including extensive local guidance.
2. Facilitate semi-automated review and evaluation of FAIRness, while retrieving information required for RDM administration
3. Conform to the RDCS by validation against maDMP-schema-1.1¹⁷

The reason for performing this pilot work in DMPOnline was based mainly on practical considerations. SU had already earlier signed up as an institution for the DMPOnline service at DCC, and even before the beginning of the EOSC Nordic project, SU were offering it as a service to their researchers and had provided instruction for PhD students on how to use it.

¹⁷

<https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/blob/df3aada9359cca14ea4a02374512a5c165c818ff/examples/JSON/JSON-schema/1.1/maDMP-schema-1.1.json>

15

So there was a perceived local need to develop this service, as RDM administrators also realized they would never be able to keep up with the growing demand for reviewing DMPs, as long as these were largely in the form of long, full text essays and there were no common criteria for how to evaluate them. Another reason for the choice of DMPonline was the later free provision of DMPonline as a test service to all Swedish universities and institutes of higher education for a limited test period. Thus we anticipated that other institutions could also benefit from the work done on DMPonline. Subsequently, during the ongoing work within the project, several external presentations of this work to representatives from other Swedish universities (Umeå, Linköping, Örebro, KTH) and also to a cross-Nordic climate research community, NICEST¹⁸ were given.

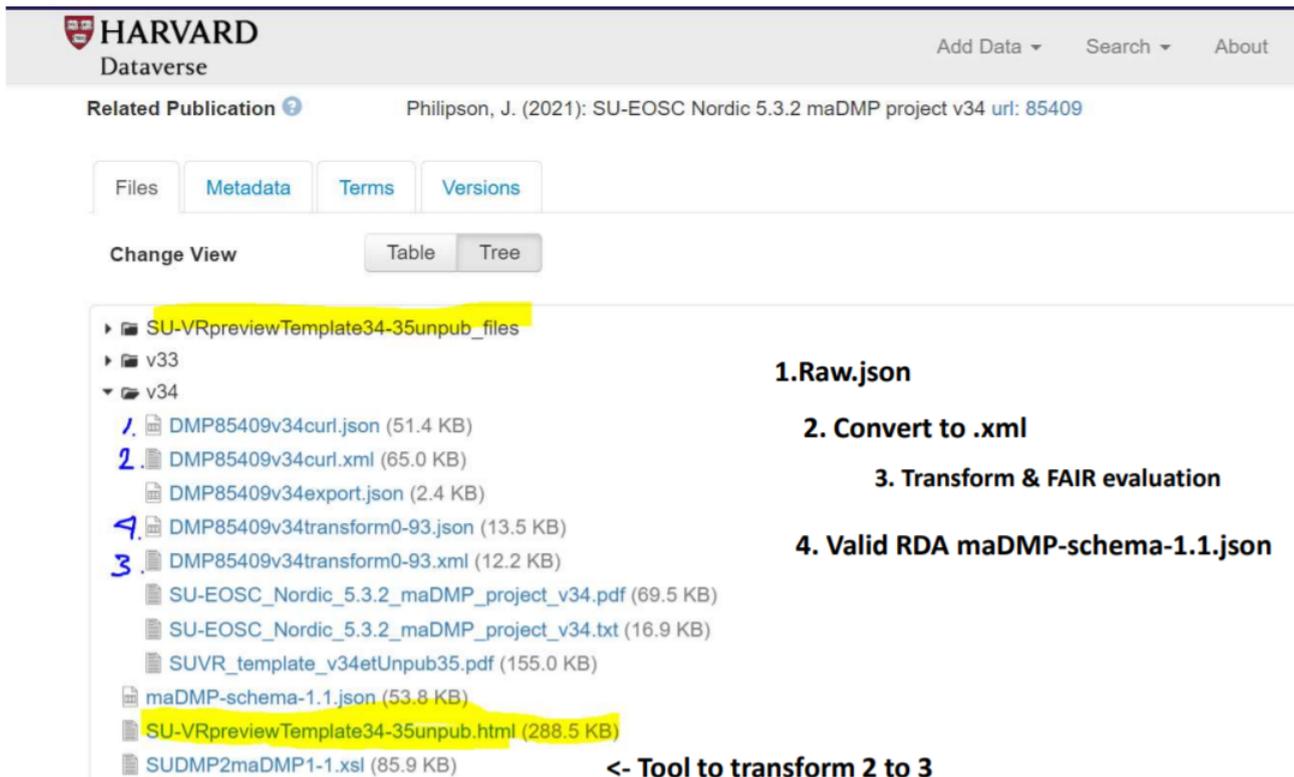


Figure. 5 Screen-shot of the SU-EOSC Nordic 5.3.2 maDMP project Dataverse draft record showing order of workflow in processing API(v0) output from DMPOnline.

The following is essentially the description of the project given in Dataverse, where a draft record and dataset for SU- part of the project can be found¹⁹ (see Figure 5):

The Stockholm University SU-VR template in DMPOnline is based on Swedish Research Council (VR) and Science Europe model (sections I-VI and original questions), but with more constrained question and answer options by means of multiple choice checkboxes, dropdown menus, radio buttons for increased machine actionability of output. The template has possible answers formatted with respect to SU Research

¹⁸ <https://neic.no/nicest2/>

¹⁹ [Private link](#), to be replaced once published with this DOI: [10.7910/DVN/MGZBAL](https://doi.org/10.7910/DVN/MGZBAL) (not yet active)

Data Policy, local research data management rules and the RDCS. The objectives are to make it easier to fill out for the individual researcher and for the output (from DMPonline API v0) to be more machine-actionable, thus facilitating review, validation against RDCS maDMP-schema and evaluation of potential FAIRness of data management measures described in the DMP. Contained in the Dataverse dataset are:

- **Version 33** of the **template** involves a change to a one phase model - Initial and Full DMP in one phase, motivated in part by an effort to avoid repeated entries of the same information. Instead, a final section *IX: Full DMP - additional Datasets and identifiers, Reference list and Project end* has been introduced to mark the completion of the research project described in the DMP.
- **Version 34** of the **template** (published in DMPonline 2021-10-05) are questions in section VIII DMP administrative information:Q5-Q7 about Funder, Grant ID and Funding status. (These are otherwise either not part of the output from custom templates like this, or possible to enter and save in the Project details description.)
- The new **version 0-93** of the **XSLT transformation** file SUDMP2maDMP1-1.xsl holding a **funderName2fundRefIDmap** then leverages the answers to VIII:Q5-7 for enhanced compliance with the RDA DMP Common Standard.

From 2021-11-02 this dataset also includes a (draft) Schematron schema for assessment of prospective FAIRness of RDM-measures described by the DMP. DMPonline instances using template (v33 and v34) with raw API (v0) JSON output are converted to XML, then transformed using the XSLT-file and converted back to JSON to check compliance with maDMP-schema-1.1. All these are included in the Dataverse dataset.

Challenges with maDMPs

The framework for constructing a maDMP template within DMPonline has its limitations. Among the challenges were:

- To get maDMP **output** was dependent on a well-functioning API (v0) from DMP Roadmap, which at times was not fully operable. API v0 gives the full content, questions and answers, except for the cover sheet with Project Details such as Funder and Grant Number (in DMPonline as a separate tab of the DMP). This is a serious limitation, since this is information that is asked for in the RDCS. Moreover, DMPs created from one of the public funder templates apparently get the Funder field in the DMP automatically pre-populated, whereas for DMPs created from local templates it currently appears impossible to fill out this field at all, not even manually. In response to this, DMPonline has mentioned the possibility of eventually “decoupling” templates and funders in the system to solve these problems by showing greater flexibility. API v1, which is still being developed and not fully implemented in DMPonline, on the other hand delivers an output of information content exclusively meant to satisfy and comply with the RDCS. This means in practice that parts of the original content may be left out, while other parts are satisfied by producing e.g. a generic, “fake” dataset entry, since this is a mandatory element in the RDCS that is still lacking in perhaps most funder templates.
- To make output (in JSON) compliant with RDCS, it first had to be converted to XML and transformed (by XSLT) and then reconverted to JSON, in order to validate against the RDCS *madmp-schema-1-1.json*.
- The conversion from JSON to XML and back to JSON cannot handle or distinguish pure numbers from strings, which are required for *dataset_id*-s and made it necessary to prefix the output from these input fields with an ‘ID-’ to ascertain getting a string datatype for validation against the RDCS *maDMP-schema-1.1*.

- The DMPonline framework does *not* allow for *pre-ingest validation of data types* or pattern matching of input strings in the template webform. This means there is little possibility of controlling or influencing the type of input / output received, other than through very clear guidance that may not always get the attention necessary to avoid less meaningful answers. We have seen some examples in the DMPs using the SU-VR template where e.g. the *dataset identifier* field was filled out with a longer description or just a simple id type such as “url” (that otherwise should be given as answer to the following question, using a drop-down menu).
- RDCS is based on the DCAT²⁰ model for datasets, allowing repeated entries for datasets and their distributions; this is not easily accommodated within the DMPonline framework for templates, and involved allowing for “dummy” dataset construction as a default in the transformation of output to comply with the RDCS.
- Originally the SU-VR template was based on a *two-phase* model, one initial phase (for complying with funder requirements), and one final phase for the end of a research project. Both phases were meant to have largely the same questions answered, only at different times, so this was intended as a kind of *versioning* of the DMPs. But since it proved difficult to copy the template web-form from one phase to another, it was decided to use a one phase model, with the addition of an extra final section to be filled out (possibly continuously during the whole project).
- Already from the beginning there was an ambition to share with others as much as possible of the work in this project, and the project has been presented in other RDM and research environments (Swedish, Nordic and European) at several occasions. (It will be presented again at the DMPonline Drop In session on October 26., 2021 at 11:30 CEST). Within DMPonline, you can share individual DMP instances with other users by invitation, and individual DMPs could also be made publicly available, but then again only as PDFs (as we have done with an earlier version of the [SU-maDMP project](#)). But the maDMP template itself, the basis of our work, can only be shared as a PDF, which is not ideal for allowing editing and adaptation to local needs at other universities. This problem has been evident also for research projects with members from different institutions or universities, even including SU, but where the lead data manager and responsible for the creation of the DMP has another, external affiliation and thus cannot access the template through dmp.su.se. To overcome this inconvenience and be able to share the template in a more flexible, editable format, various PDF conversion tools have been evaluated. However, at least one of them appears to give a rather distorted rendering of the template structure. Another way of sharing that we tested recently is to provide a preview-copy of the latest SU-VR template version in HTML format, that could then be viewed and edited in accordance with local needs. This preview-copy of the template (v34-35unpub) as HTML is now available for download from the Dataverse draft record through the [private link](#) referred to above. However, it will still not be possible to create a local template in DMPonline from this rendered preview alone, without using copy - paste.
- Local templates (such as the SU-VR template), as distinct from public funder templates, will not get the Funder and Grant Number information from the cover sheet (Project Details tab) as part of the output. Moreover, DMPs created with such templates will not even have the possibility of filling out and adding Funder information, which is needed for full compliance with the RDCS.
- The many different versions and updates of the template model, generating different real DMP instances with, in some cases, a different order of questions and answers, has necessitated a continuous update and adaptation of the XSLT-file to accommodate changes. This is required so that the output of DMPs using the SU-VR template, independent of version, can still be transformed in accordance with the requirements for validation against the RDCS maDMP-schema-1.1.json, and using the same FAIRness assessment metrics for all versions.

²⁰ <https://www.w3.org/TR/vocab-dcat-3/>

We encountered further challenges that were not directly dependent on the DMPonline framework, but rather on the RDCS and different FAIR metrics tools:

- Assessment of (potential) FAIRness of RDM measures in the research project, described by the DMP, is not part of the RDCS proper, and constitutes an extension.
- FAIR metrics schemes developed so far, e.g. the FAIRsFAIR v0.4 (doi: [10.5281/zenodo.4081213](https://doi.org/10.5281/zenodo.4081213)) used by the [F-UJI tool](#), are more adapted to evaluation of the machine-actionability structure of general data *repositories*, rather than individual datasets or RDM measures in research projects.

Potential benefits compared to other approaches

There are some advantages to an approach based on a local template instead of trying to find a smallest common denominator in general funder templates in order to achieve compliance with the RDCS, as e.g. DMP Roadmap / DMPTool and tools like Argos seem to aim for.

- Possibility to tailor output of DMPs to local administrative information needs and policies, allowing for at least semi-automated FAIR-assessment and potential self-evaluation. In this way responding in part to the well-founded criticism by *Smale et al.* [7] of DMPs hitherto not achieving their stated mission of serving the needs of different stakeholders.
- Less need to make up dummy entries of input fields (e.g. datasets and identifiers) that are required output in the RDCS, but are not commonly part of general funder templates.
- While our local model, using DMPRoadmap/ DMPonline API v0, conversion to XML, transformation with XSLT and conversion back to JSON actually produces RDCS compliant output that validates fully against the maDMP-schema-1.1, the current export to JSON within DMPonline, using instead API v1 does not yet fully validate against the maDMP-schema-1.1. Required keys, such as contact and contributor_id are missing and the data formats used are incorrect, even when the DMPonline DCC default template is used for the creation of a DMP.

Recent and future development

In the DMPonline user group meeting 2021-11-02, we learned that some of the features of our SU-VR template may become obsolete (or redundant) if DMPonline will also in the near future implement the DMPRoadmap features that are already found in DMPTool (at CDL - California Digital Library). These features include elaborate versioning of DMPs, integration with ORCID²¹ and a particular tab for Research Outputs with repeatable fields for datasets, metadata standards and licenses, as a foundation for compliance with the RDCS. With that in mind, further development in the SU-EOSC Nordic 5.3.2 maDMP project will focus mainly on the FAIRness assessment measures and the Schematron schema.

5.1.2 Machine actionable plans in EasyDMP

The work attempts to address three challenges:

1. Can we adapt the tool to support export and import of maDMPs, in particular the RDCS? This would allow integration with services such as the NIRD storage allocation service (MAS).
2. Can we enable interoperability with the DSW for the Elixir-No community? This would allow plans to be transferred between tools reducing the need for researchers to fill in the same information more than once.
3. Can we support interoperability between EasyDMP and the NIRD archive? This would reduce the need for the researcher to fill in some of the metadata information required for archived datasets.

²¹ <https://orcid.org/>

In the subsections below we describe the current status of each of these challenges.

Supporting RDCS

EasyDMP is an existing web-based tool for making DMPs. It was started as a reaction to existing systems that needed domain specific knowledge on how to write an acceptable DMP, in an essay format. Instead of an essay, EasyDMP uses a wizard/make your own adventure format where the researchers choose among a small set of answers. The answers can be converted to an essay, by inserting the answers into a frame as shown in Table 1 below.

Question	How much space do you anticipate you will need to archive the research data?
Answer type	An integer representing terabytes, with full input validation
Frame	“The research data is estimated to need about BLANKETY BLANK terabytes of storage.”
Specific answer in a specific dmp	10
Resulting text in the essay version	“The research data is estimated to need about 10 terabytes of storage.”

Table 1. An example of a question structure. The researcher fills in the answer and EasyDMP uses the Frame to create a response.

In EasyDMP, a collection of questions, frames and their allowable answers is called a *template*. A template is divided into one or more sections, and each section contains questions, their answer-alternatives according to answer type, and their frames. When a researcher has answered all the required questions of a template, the result is a *plan*, which is available both as an essay, and as JSON-structure of question-answer pairs.

The actual answers are stored in a machine-*readable* format so that they can easily be edited and a new essay generated. Since the original use case was mass-generating essays according to funders’ needs, the answers are not machine-*actionable* however, because the question/answer pairs do not carry any semantics/metadata.

While EasyDMP and similar funder-driven systems are organized in a linear, very open structure, completely in control of the template designer, the RDCS is structured around *objects* containing other objects that are not designed to be converted into free-flowing text. According to the standard, a plan contains at least one contact, one or more datasets, zero or more contributors, zero or more projects, and zero or more cost-objects. Each project-object may contain zero or more founding objects, each dataset may contain zero or more distribution-objects etc.

In order for EasyDMP to be able to interact with plans built according to RDCS, it was necessary to adapt it to support the objects-within-objects structure. This was done by adapting the *sections*: these were changed to be able to be optional, to be able to contain other sections, and to be answered more than once. This stage of the adaptation is complete.

The next stage is to design a complete template following RDCS, and build a system to import and export plans using this template. Once there exists one such complete plan, the framing-system can be adjusted so that an RDCS plan can be turned into an essay format. When this stage is finished we hope new templates will use the RDCS template as a basis so that the standard is followed by default.

The final stage is to add semantics to the question/answer pairs, and map or rewrite existing non-RDCS templates to be RDCS compliant. This will have to be done manually per template and we plan to start with the Science Europe template.

Our work to date on implementing the RDCS resulted in a contribution to the paper [3] on machine actionable DMPs.

Interoperability with the DSW and NIRD services

The ELIXIR-NO community has embarked on a project to integrate with the NIRD infrastructure. The goal is for Norwegian ELIXIR-NO researchers to be able to seamlessly use NIRD resources without having to explicitly request those resources. The necessary information required by the NIRD services would be extracted from the DMP (See Figure 6).

Proposals for NIRD resources require a DMP to be submitted with the proposal. Researchers use EasyDMP to create their DMPs as it has a DMP template tuned to the NIRD services (questions match the service requirements). Once the work on integration of EasyDMP with NIRD services has completed (aiming for the end of next year for some services) the services will be able to automatically consume the DMP information instead of manually as is the case now.

However, ELIXIR-NO researchers use the DSW to create their DMPs and it would be desirable if the researchers did not have to fill in the same information in the EasyDMP system. So, our first task is to allow a researcher's DMP to be exported from the DSW and imported into EasyDMP as shown in Figure 6.

Use case 1



Figure 6. Illustrates the ELIXIR-NO use case. A researcher on the left creates a DMP in the DSW. The plan is exported and imported into EasyDMP and the information is then consumed by NIRD services.

A first step for this task was the examination of how the RDCS could be used as the schema to transfer the plan from DSW to EasyDMP. Our investigation indicated that the RDCS is able to support approximately 4% of the full ELIXIR-NO knowledge model (in any one plan not all of the knowledge model is used, so the coverage will actually vary). The RDCS was also compared with the EasyDMP template and was found to cover approximately 45% of the EasyDMP template. Conversely, the EasyDMP template covers

approximately 50% of the RDCS. However, we found that a question in the EasyDMP template mapped to more than one element in the RDCS. So the actual coverage is more difficult to determine as it depends on how difficult it is to extract the information for each element. We expect further analysis of the two schemas to improve the coverage (this is mainly due to developing a better understanding of the elements of the RDCS which would reduce the ambiguity on some questions).

This result was a little disappointing as it shows that there will be a considerable loss of information when exporting the DMP in RDCS format from the DSW and importing into EasyDMP. Nevertheless, we are interested to see just how bad the loss is in practice.

The DSW includes a first implementation of a JSON export that is compliant with the RDCS schema and we are currently working on the API to support RDCS import into EasyDMP (as described in the previous section). At the same time, we are manually importing the exported DMP into EasyDMP in order to exercise a part of the integration with a NIRD service (namely the NIRD Research Data Archive). This will also give us quantitative data regarding how much information loss is occurring. The next step will be to make use of the API to automatically export a DSW DMP and import it into EasyDMP.

Since we already know that information will be lost we are working on an extension to the RDCS schema where we include the missing information. The schema will need to be agreed between the DSW and EasyDMP and it will only be complete for the chosen Knowledge model and template. This pragmatic approach will work for our needs (we expect changes to either DMP schema to be very limited), but it does not seem to be a satisfactory solution for much wider use.

Once we have an agreed schema we will create new versions of the APIs and test the export and import. Once we are happy with the results, the system will be used for subsequent requests for resources. We anticipate we will have 60-70% of the use-case complete (i.e. export of the DMP and automatic import into EasyDMP) towards the end of next year with the completion of interoperation of EasyDMP with NIRD services by summer the following year.

5.1.3 Machine actionable SND checklist

The latest version of the SND checklist described in Section 4.4 was released very recently. The resulting plans are still documents that are not machine actionable. We intend to map the latest version of the checklist to the DMP tools (easyDMP and DSW) over the remainder of EOSC-Nordic which will allow us to export RDCS compliant plans.

This proof of concept will be useful at a later stage because in May 2020, SND launched a cohesive system which makes it easier for researchers to disseminate data through SND. The tool, called DORIS (the SND Data Organisation and Information System), is one of several large-scale development projects in which SND collaborates with researchers and staff in the local data support units (DAU) in Swedish higher education institutions. This tool can be used throughout the process of making research data accessible: from describing research data to ordering data for re-use or review. The development of the system has been divided into several phases. Within DORIS, some of the information is similar to that found in DMPs and it is connected to the universities where it harvests administrative information. It is possible to export the information back to the universities.

One possible and in many ways wanted solution among researchers and other stakeholders is that there is ONE system where researchers can produce funding applications, ethics applications, DMPs, and where they can make their data available for reuse. That would be a single tool that the researcher and the other stakeholders can use, that can harvest information from different parts of the university systems etc. and then return information back to those systems. It is essential that the information in this system would be

protected/stored and locked/shareable if and when the researcher wants it to be. This tool cannot have its information totally machine actionable, but has to be human readable at least for some sections (after all, the information is supposed to be used in a web catalogue as well). This tool also has to have the capabilities to produce/export versioned DMPs, funding applications, ethical applications, etc. A first step to create a tool like this would be to map SNDs checklist to EasyDMP and/or DSW. Depending on the result, the experience can then be used to implement a DMP tool into the DORIS system.

6. Discussion

Our work on enabling maDMPs by implementing the RDCS has provided us with a new viewpoint on data management plans that we think is worth mentioning. We also describe the lessons we have learned in implementing and using the RDCS.

6.1 The DMP Lifecycle

Implementing maDMPs that can be consumed by other services has made us realise that the life-cycle of the DMP needs to be clearly defined. The service that consumes the DMP needs to know when it can access the DMP. The DMP needs to have well defined states that a machine can recognise and it needs to have the concept of a version. Otherwise, a service consuming the plan would be uncertain if any changes it sees in the plan are deliberate or an accident. Plan versions would need to depend on approval, either by an external body, or by the project itself.

Ideally, DMPs would need a unique, persistent identifier, but when this identifier is issued is not so clear. If the DMP is considered as a document that is living then we have to think how to accommodate this. This problem is quite similar to the problem of researchers collecting sensor or continuous data. In those cases, versions or editions are created each with their own DOI and the metadata keeps track of which DOI belongs to which version. In the case of DMPs each DMP would need to keep track of any daughter or parent DOIs to provide a link to different versions.

Another issue is that DOIs need to point to a specific location. The DMP would need to contain a copy of its own DOI otherwise information on where the DMP came from is lost and there is the potential for confusion as one researcher may notice an error in a Zenodo copy of a plan, download it, edit it and upload it to a different repository and get a new DOI. To a third researcher, it would be unclear which DMP is more valid.

6.2 Lessons Learned

One point that has become very clear to us throughout the course of our work is that current DMPs start from the viewpoint of the project and describe the environment in which the datasets are created. They are a plan in the truest sense as they express the intention of how the data will be managed. The RDCS DMP takes the viewpoint that the dataset already exists and the dataset is central to the plan as shown in Figure 6. It is not really a plan, but more a document of what was done.

Since these two view-points are different, it is quite natural that there is a degree of incompatibility between the two. Both are relevant. The plan is necessary in order to understand beforehand what services and resources are needed and the *Data Management Document* (represented by the RDCS) is necessary for researchers that wish to reuse the data, or for administrators that wish to know, for example, how many projects used the institution archive to store their data, or other consumers.

The problem we are facing is how to translate between these two different viewpoints. One approach described in Section 5.1.1 is to use dummy values for entities that cannot be determined by the plan (such as the dataset) in order to create a RDCS compliant DMP. Although, one has to be careful to replace the

dummy entities with the actual values once they are known and then make sure the correct dataset references the correct metadata. Another workaround is to extend the RDCS with elements that correspond to the information. But, this is brittle as it results in a mapping between a particular classic DMP template and the RDCS. If the classic template releases a new version with changes to existing questions the whole mapping needs to be redone (unless sections are isolated).

A perhaps more viable variant of the approach just described is a structure of one machine-actionable part and one textual part. This seems to be the path chosen recently by DMPTool (CDL) within DMPRoadmap, offering a special tab for Research Outputs connected to the DMP, serving as basis for the export to and compliance with the RDCS.

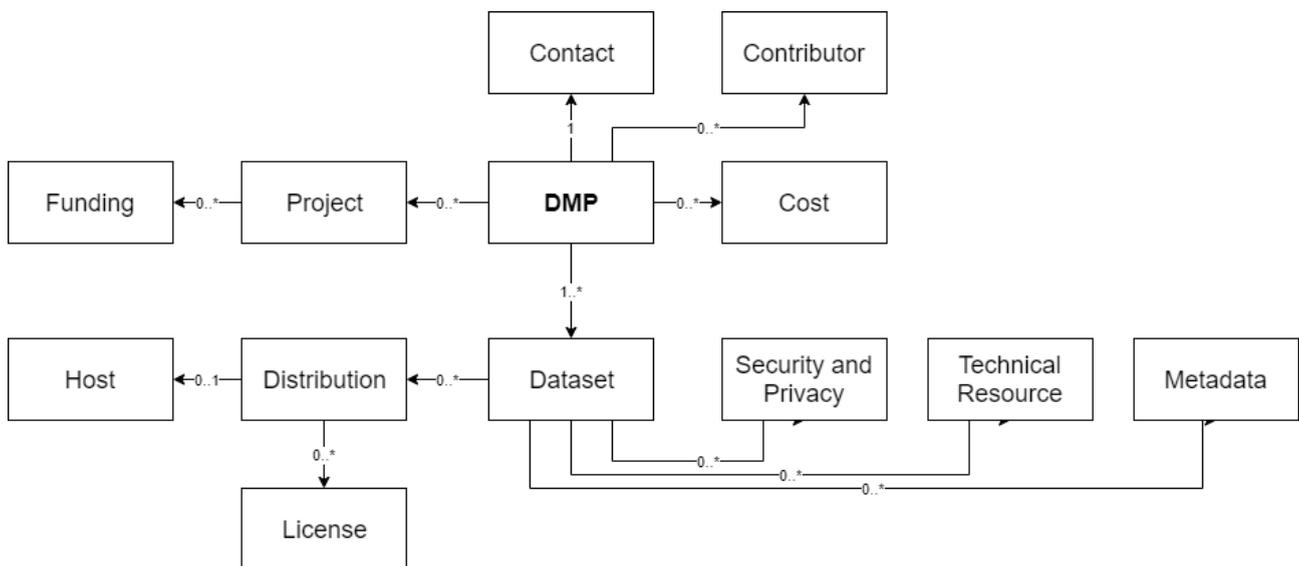


Figure 6. The RDCS application profile for a DMP taken from [5]. The DMP box only contains high-level information on the DMP such as title, description.

Another option is to simply implement templates that follow the RDCS. This is possible for people creating new plans, but not for people with existing plans (of which there are many). Also, we have seen researchers tend to create a new DMP by copying an existing DMP and modifying the relevant parts. Such a change to RDCS would require researchers to map existing plans to the new schema.

Many of the DMP tools are just starting to provide APIs that support the RDCS schema, so there is naturally a degree of change as the APIs mature which makes integration with other services difficult.

Another lesson we have learned is that one guideline on data management (such as Science Europe) can be interpreted in more than one way resulting in many DMP templates that are subtly different. The reason for this is quite clear: the guidelines are not precise, so there is a latitude of interpretation. This makes interoperation between tools difficult. We should mention that Science Europe does offer a rubric²² for evaluating DMPs which can help to understand what a guideline is expecting. A much better approach would be for Science Europe to offer a template that would implement the basic questions that result in a plan that conforms with the guidelines. Institutions would then be able to use the template as a basis and implement on top of that.

²² https://surveys.scienceeurope.org/upload/surveys/228819/files/20200703_rubric.pdf

Each template should ideally have an ontology to make mapping between templates easier. This would make importing and exporting plans from one tool to another where different templates have been used simpler.

Another observation is that the services that could consume information from a DMP currently request the information to be supplied to the service directly. For example, the NIRD archive requires the researcher to input metadata into a web form. Consuming information from a DMP would require the archive to create a service that could pull information from the DMP and populate the appropriate metadata fields. In some cases (such as the NIRD archive) this is not difficult. It requires the development of an additional service that can populate the metadata catalogue directly. However, some services or broker services are black boxes where the metadata input is tightly coupled to the metadata store making it very difficult to supply metadata via a different mechanism. In these cases it is not so clear how to proceed.

6.2 The wider context

In a wider context the relevance of DMPs on improving FAIR data is being recognised.

Funders of science in a growing number of countries are requiring researchers to use DMPs at the start of projects and are encouraging researchers to make their datasets FAIR. As a result, we see in more countries that funders are requiring maDMPs (as they can capture how the data was managed which helps to increase the reuse of data) and several funders are considering financial stimuli to drive this FAIR behavior.

In early September 2019 a number of funders met in Leiden^{23 24} to discuss the necessary structure to drive researchers and data stewards towards producing appropriate Data Management Plans, using generic and domain specific metadata templates for increased interoperability.

Earlier this year 2021, the Research on Research Institute (RORI²⁵) consisting of about 20 associated funders and partners defined a project to develop a supporting tool for researchers called FAIRWARE²⁶, based upon the FAIR Principles, DMP's and Metadata Templates, using Metadata for Machine (M4M) workshops as a mechanism to define the required metadata for the maDMP's. For this development, RORI has teamed up with the University of Stanford's Biomedical Informatics Research (BMIR).

It is foreseen that funders in the future may **insist** on FAIRness as a condition for approving grant-applications, thereby offering financial stimuli for researchers.

7. Next steps

In this document we have described our work on enabling maDMPs by implementing the RDCS in order to support our stakeholders. In EOSC-Nordic our plan is to continue to complete our proof of concepts and integrations. Some of our work will extend beyond the end of the project and these are already in our institution roadmaps as the integration of DMPs with services is considered an important task.

²³ <https://www.go-fair.org/events/fair-funders-meeting/>

²⁴ <https://www.go-fair.org/2019/11/04/fair-funder-implementation-study/>

²⁵ <https://researchonresearch.org>

²⁶

<https://www.eoscsecretariat.eu/eosc-liaison-platform/post/rori-funders-consortium-selects-bmir-fairware-project>

25

We foresee that in the not too distant future maDMPs, with explicit considerations for the creation of Open, FAIR and domain-specific research outputs, will become indispensable roadmaps for navigating the research data life cycle.

We foresee that the maDMP will coordinate the respective roles of funders and research institutes (who will as policy, mandate more and more such plans), guiding the execution of projects by researchers and data stewards, and assist the dissemination of their findings by publishers (be they traditional research articles or datasets). The routine use of maDMPs will create immediate and visible benefits to each of these stakeholders, saving time and money in relocating and reusing data, code and entire workflows. At macro level this may lead to significant monetary benefits in research worldwide.

We foresee that the constellation of decisions regarding data FAIRification, when captured as FAIR Implementation Profiles (FIP's), will be directly coupled to maDMPs. Not only does this relieve the researcher of the burden of crafting FIPs and DMPs each time from scratch, but it provides a reference for industries and domains to continue to design, build and perfect their agreed vocabularies and metadata templates enabling a higher level of semantic interoperability within and between domains. Researchers and data stewards will see increasing efficiencies by reusing high quality and domain relevant MaDMP's created by their fellow researchers.

The timing of turning this vision into reality is not a function of today's existing technology, but depends strongly on the willingness of the different stakeholders to cooperate in order to make the delivery of a maDMP a standard exercise in research.

8. References

- [1] Gajbe S. B., Tiwari, A., Gopalji, Singh, R. K. (2021) "Evaluation and analysis of Data Management Plan tools: A parametric approach", *Information Processing & Management*, 58(3), pp. 10248. doi: [10.1016/j.ipm.2020.102480](https://doi.org/10.1016/j.ipm.2020.102480).
- [2] Kim S. (2020) "Machine-actionable Data Management Plans Model Analysis and Improvement Direction," *Journal of Information Science Theory and Practice*. Korea Institute of Science and Technology Information, 8(4), pp. 20–28. doi: [10.1633/JISTAP.2020.8.4.2](https://doi.org/10.1633/JISTAP.2020.8.4.2).
- [3] Miksa, T., Walk, P., Neish, P., Oblasser, S., Murray, H., Renner, T., Jacquemot-Perbal, M.-C., Cardoso, J., Kvamme, T., Praetzellis, M., Suchánek, M., Hooft, R., Faure, B., Moa, H., Hasan, A. and Jones, S., 2021. Application Profile for Machine-Actionable Data Management Plans. *Data Science Journal*, 20(1), p.32. DOI: <http://doi.org/10.5334/dsj-2021-032>
- [4] Miksa T, Simms S, Mietchen D, Jones S (2019) "Ten principles for machine-actionable data management plans." *PLoS Comput Biol* 15(3): e1006750. <https://doi.org/10.1371/journal.pcbi.1006750>
- [5] Miksa, Tomasz, Walk, Paul, & Neish, Peter. (2020). "RDA DMP Common Standard for Machine-actionable Data Management Plans." <https://doi.org/10.15497/rda00039>
- [6] Simms S., Jones S., Mietchen D., Miksa T. (2017) "Machine-actionable data management plans (maDMPs)." *Research Ideas and Outcomes* 3: e13086. doi: [10.3897/rio.3.e13086](https://doi.org/10.3897/rio.3.e13086)
- [7] Smale, N. et al. (2020) "A Review of the History, Advocacy and Efficacy of Data Management Plans ", *International Journal of Digital Curation* 2020, Vol. 15, Iss. 1, pp.30 . doi: [10.2218/ijdc.v15i1.525](https://doi.org/10.2218/ijdc.v15i1.525)

Quote

“The desirability of not having a zillion templates”
