Application Profile for Machine-Actionable Data Management Plans

TOMASZ MIKSA ⓘ

PAUL WALK ⓘ

PETER NEISH ⓘ

SIMON OBLASSER ⓘ

HOLLYDAWN MURRAY ⓘ

TOM RENNER ⓘ

MARIE-CHRISTINE JACQUEMOT-PERBAL ⓘ

JOÃO CARDOSO ⓘ

TROND KVAMME ⓘ

MARIA PRAETZELLIS ⓘ

MAREK SUCHÁNEK ⓘ

ROB HOOFT ⓘ

BENJAMIN FAURE ⓘ

HANNE MOA ⓘ

ADIL HASAN ⓘ

SARAH JONES ⓘ

Author affiliations can be found in the back matter of this article

ABSTRACT

This paper presents the application profile for machine-actionable data management plans that allows information from traditional data management plans to be expressed in a machine-actionable way. We describe the methodology and research conducted to define the application profile. We also discuss design decisions made during its development and present systems which have adopted it. The application profile was developed in an open and consensus-driven manner within the DMP Common Standards Working Group of the Research Data Alliance and is its official recommendation.

1 INTRODUCTION

Data Management Plans (DMPs) are documents that accompany research proposals and project outputs. "They describe the data that is used and produced during the course of research activities, where the data will be archived, which licenses and constraints apply, and to whom credit should be given" (Miksa, Simms, Mietchen and Jones [2019]). The existing practice of writing DMPs is primarily driven by research funders who consider DMPs not only to be planning, but also a steering and evaluation tool. However, DMPs are often perceived by researchers as an annoying administrative exercise that does not support data management activities (Smale et al. [2018]). This is because answering specific questions requires special knowledge or the information requested overlaps with data previously submitted elsewhere. This in turn has impact on the quality of DMPs, because questions remain unanswered, or the free text answers are copied between DMPs for different projects, and lack specific details.

DMPs should continue to be presented in a human-readable way. However, this information must be complimented by a machine-actionable representation consisting of atomised, structural data. Breaking the information down into specific fields creates added value to all stakeholders in the research data lifecycle such as researchers and funders, but also data stewards, repository operators, etc. The added value is created when the data is automatically collected as well as re-used by systems acting on behalf of stakeholders. An example would be a repository setting information on backup strategy and preservation policy in response to a data steward choosing that particular repository for data deposit. Similarly, information from DMPs can be used to trigger actions, for example, the license and embargo selected by a researcher can be used to automatically fill out information on data deposited into a repository.

To realise the vision of machine-actionable DMPs (maDMPs) as a way to exchange and act on information about data used and produced by researchers, we need all stakeholders to collaborate and synchronise their efforts. The basic framework requires: (i) an application profile for representing information in a common way; (ii) services that can provide and use this information in an automated way. The application profile is the focus of this paper and can be defined as a metadata design specification that uses a selection of terms from multiple metadata vocabularies, with added constraints, to meet application-specific requirements.[1]

The Research Data Alliance (RDA) recognised the importance of making DMPs machine-actionable and established the DMP Common Standards[2] working group to develop an application profile allowing for "automatic exchange, integration, and validation of information provided in DMPs and facilitating the exchange of information between systems acting on behalf of stakeholders involved in the research life cycle" (Miksa, Walk and Neish [2019]).

In this paper, we describe the application profile for machine-actionable DMPs developed by the DMP Common Standards working group. We present the research conducted and methodology used to define it. We describe the motivation and rationale behind design decisions made. We also present a range of services adopting the application profile as examples of how it can be used to create value for stakeholders. This is the first paper describing in a holistic way all the work done to release the official version of the application profile for machine actionable DMPs (Miksa, Walk and Neish [2019]), as well as its adoptions.

The following sections of this paper are structured as follows. In Section 2 we describe the related work. Section 3 describes how the application profile was developed. Section 4 describes the application profile using an example of a minimal maDMP. Section 5 provides discussion on the application profile, design decisions made, etc. Section 6 presents adoptions of the application profile. Section 7 presents conclusions and future work.

2 RELATED WORK

"Data management plans are required by funding bodies and institutions all over the world, e.g. the National Science Foundation (NSF) in the USA, the European Commission in Europe, or the National Research Foundation (NRF) in South Africa" (Miksa, Neish, Walk and Rauber [2018]).

---

1   https://www.dublincore.org/resources/glossary/application_profile/.

2   https://www.rd-alliance.org/groups/dmp-common-standards-wg.

Researchers are often advised to follow the principles defined in Michener [2015] to write a good and comprehensive DMP. There is also a wide range of tools supporting researchers in the creation of a DMP, e.g. DMPonline,[3] DMP Tool,[4] Data Stewardship Wizzard[5] or RDM Organizer (Engelhardt et al. [2017]). These tools provide questionnaires that must be answered to create a DMP complaint with a selected funder template. Jones et al. [2020] present a table describing most recent DMP tools.

Science Europe, which brings together research funders from Europe, issued common guidelines for development of DMP templates (Doorn [2018]). Hence, many templates used across Europe are similar. Furthermore, they are also often derived from the checklist created by the DCC (DCC [2013]).

Machine-actionability (defined as "information that is structured in a consistent way so that machines, or computers, can be programmed against the structure")[6] and DMPs are of interest to the Research Data Alliance[7] (RDA). The RDA facilitates the formation of community-driven groups to deal with topics related to data management and sharing. One of them is the the RDA DMP Common Standards[8] working group that developed the application profile described in this paper. Community needs that led to establishing this working group are described in the paper by Simms et al. [2017]. An initial model for maDMPs and its mapping to tools and standards was presented in Miksa et al. [2017] and was used to kick-off developments by the group.

We use the term application profile as defined by Dublin Core:[9] "An application profile is a metadata design specification that uses a selection of terms from multiple metadata vocabularies, with added constraints, to meet application-specific requirements". This definition emphasises that our goal was to reuse existing standards, models, terms, vocabularies, etc. Other examples of application profiles include the European Commission who introduced an application profile for data portals in Europe[10] that complements the W3C DCAT standard (Archer [2014]) and describes which fields in W3C DCAT are mandatory and must be used by all repositories adopting the application profile. In a similar fashion, RIOXX application profile[11] was designed to fulfil requirements of Research Councils UK (RCUK) for institutional repositories to share metadata about the scholarly resources.

3 METHODOLOGY

In this section we present how the application profile was developed within the DMP Common Standards working group. We describe the open stakeholder consultation process, as well as tools and accompanying concepts developed.

We followed the Design Science research methodology (Hevner et al. [2004]), which is common in computer science, especially in information and knowledge-based systems. Hence, given the problem description, we iteratively produced artifacts (methods, models, prototypes) and evaluated them accordingly.

All these activities helped in defining what machine-actionable DMPs are and resulted in the formulation of the application profile for maDMPs.

3.1 OPEN STAKEHOLDER CONSULTATIONS

We performed two open stakeholder consultations. Each of them included both physical workshops and virtual meetings to collect feedback. The first consultation was aimed at

---

3   *https://dmponline.dcc.ac.uk*.

4   *https://dmptool.org*.

5   *https://ds-wizard.org*.

6   *http://www.ddialliance.org/taxonomy/term/198*.

7   *https://www.rd-alliance.org*.

8   *https://www.rd-alliance.org/groups/dmp-common-standards-wg*.

9   *https://www.dublincore.org/resources/glossary/application_profile/*.

10   *https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe*.

11   *https://www.rioxx.net/about/*.

defining the scope of the application profile. We applied requirements engineering methods (Dalpiaz and Brinkkemper [2018]) known in software engineering and used GitHub for collecting requirements in form of user stories. Everyone, regardless of their geographical location and role in the research data lifecycle, were able to submit their own requirements that should be taken into account by maDMPs. In total, we collected 108 user stories. The user stories involved viewpoints of funders, institutions, repository operators, research support, researchers, and service providers. The number of collected requirements and their diversity showed that we managed to engage with a wide range of stakeholders. This in turn helped in establishing a common definition of machine-actionable DMP that reflects expectations of the community. The first consultation is described in detail in Miksa, Neish, Walk and Rauber [2018].

In the second consultation we mapped the captured requirements to specific fields in existing standards and models to identify which of them can become a basis for the application profile. The second consultation was conducted together with domain experts and is described in detail in Miksa, Cardoso and Borbinha [2018]. *Table 1* presents a list of standards that we found relevant.

RELEVANT STANDARD	URI
Access License and Indicators	*http://www.niso.org/schemas/ali/1.0/*
Dublin Core Element Set	*http://purl.org/dc/elements/1.1/*
DCMI Metadata Terms	*http://purl.org/dc/terms/*
Friend of a Friend (FOAF)	*http://xmlns.com/foaf/0.1/*
DCAT	*https://www.w3.org/TR/vocab-dcat/*
DataCite	*https://schema.datacite.org*
cerif	*https://www.eurocris.org/ontologies/cerif/1.3/*
COAR	*http://vocabularies.coar-repositories.org/*
ISO 6391-1	*https://www.iso.org/iso-639-language-codes.html*
ISO 4217	*https://www.iso.org/iso-4217-currency-codes.html*

Table 1 List of relevant standards considered when developing the application profile.

We also collaborated with other RDA groups working on related topics, such as Exposing DMPs[12] and Active DMPs,[13] to collect feedback and to incorporate feedback from their consultations (Simms et al. [2018]).

The open stakeholder consultations allowed us to iteratively define the scope of the application profile and to identify existing standards that are relevant. Thus, we were able to derive an initial list of fields that should be contained in the application profile to reflect the identified needs of stakeholders.

3.2 PROOF OF CONCEPT TOOLS

Together with students of Data Stewardship at the TU Wien we developed proof of concept tools that demonstrate how existing data management practices can be improved and what new opportunities are created when maDMPs are in place. This helped us two-fold: (1) to better explain the novelty and benefits of maDMPs to stakeholders and thus get further feedback from them, (2) to further refine the application profile and to better reflect the needs of automated data processing, e.g. to group information into specific classes and to design the most suitable structure of the application profile. The proof of concept tools can be broken down into different categories:

- Tools showing how information from existing systems can be used for estimations and recommendations in an early phase of a project, e.g. for cost estimation, repository recommendation, or license selection. (Miksa, Cardoso and Borbinha [2018]).

- Tools exporting information from existing DMP tools into maDMPs, e.g. work by: Pichler [2019], Breitenfellner [2020], Inschlag and Drechsel [2020]. To do that, students

---

12 *https://www.rd-alliance.org/groups/exposing-data-management-plans-wg.*

13 *https://www.rd-alliance.org/groups/active-data-management-plans.html.*

developed simple mappings between the application profile and funder templates (EC Horizon 2020 and Science Europe). Thus, we identified that not all of the information currently requested by funders is machine-actionable. For example, the description of quality assurance processes will always be in free-text and the application profile must accommodate such non-machine-actionable information as well.

- Tools making information contained in maDMPs human-readable, e.g. work by Aigner [2020], Alkhatib and Rivera [2020], Leidinger [2020]. In other words, the tools use information from maDMPs to pre-fill existing DMP templates. Thus, the researchers do not start with an empty page when writing a DMP and funders still can receive a PDF version of a DMP.

- Tools integrating systems using maDMPs to automate upload of files into repositories by extracting information from maDMPs. The tools also support the opposite scenario: a dataset already exists in a repository and information about it needs to be added to the maDMP. The prototypes were developed for Dataverse (Hido1994 and Alhirthani [2020]), Invenio (Tsepelakis [2020]) and content management system Alfresco (Bakos et al. [2018]). We also performed analysis on how maDMPs can help in automating repository recommendation (Oblasser et al. [2020]).

Finally, we developed mock-ups (Oblasser [2020]) for the next generation tool for DMPs that puts machine-actionability at its core. The mock-ups are a result of broad stakeholder consultations performed face to face with researchers in different domains and also of feedback collected online. The mock-ups were converted into a prototype tool (Oblasser [2019]) allowing for export of maDMPs using the application profile for maDMPs.

3.3 PROCESSES AND GUIDELINES

We used the experience collected by developing and testing prototypes to define typical processes in which maDMPs can be used (Oblasser and Miksa [2019]). The set of processes is independent of any implementation and can be used by any institution or organisation as guidance on how to build the ecosystem of services that exchange information using maDMPs. The processes are not implementation blueprints and require customisation. However, they help in understanding which organisational and technical components already exist and which of them must be introduced. For example, institutions willing to provide storage to researchers must also consider cost models used for calculating costs of storage. Machine-actionable DMPs can carry information on costs, but the services of organisation must provide the actual values. The maDMP itself does not contain any business logic.

While the processes help individual organisations to streamline the discussion on machine-actionable DMPs, we also developed 10 principles for machine-actionable DMPs that call for coordinated effort within the broad research data management community (Miksa, Simms, Mietchen and Jones [2019]). The principles "contain specific actions that various stakeholders are already undertaking or should undertake in order to work together across research communities to achieve the larger aims of the principles themselves" (Miksa, Simms, Mietchen and Jones [2019]). The principles also "describe existing initiatives to highlight how much progress has already been made toward achieving the goals of maDMPs as well as a call to action for those who wish to get involved" (Miksa, Simms, Mietchen and Jones [2019]).

This helped us to better separate the concerns and narrow down the focus of the application profile, which is the information carrier and the actual automation is the task of systems using it.

4 APPLICATION PROFILE

In this section we outline the structure of the application profile and illustrate this with an example of a minimal maDMP. The full description of concepts used, as well as clarification on its purpose and usage, can be found in the official recommendation (Miksa, Walk and Neish [2019]) and in its official repository.[14]

---

14   *https://github.com/RDA-DMP-Common/RDA-DMP-Common-applicationprofile*.

Figure 1 presents concepts used within the application profile. Each concept is further broken down into specific fields (not depicted). The central concept is the *DMP* which provides generic information on the DMP. It is used to link information on projects, costs, contributors and contact point for the DMP. Most of the relations are optional (0.*) and depend on the specific setting in which the maDMPs are used. Each *DMP* must have at least one *Dataset*. Datasets are used to group information on the data described by a DMP. This includes information such as: size, format, location of the data, existence of embargo, metadata standards used, etc.
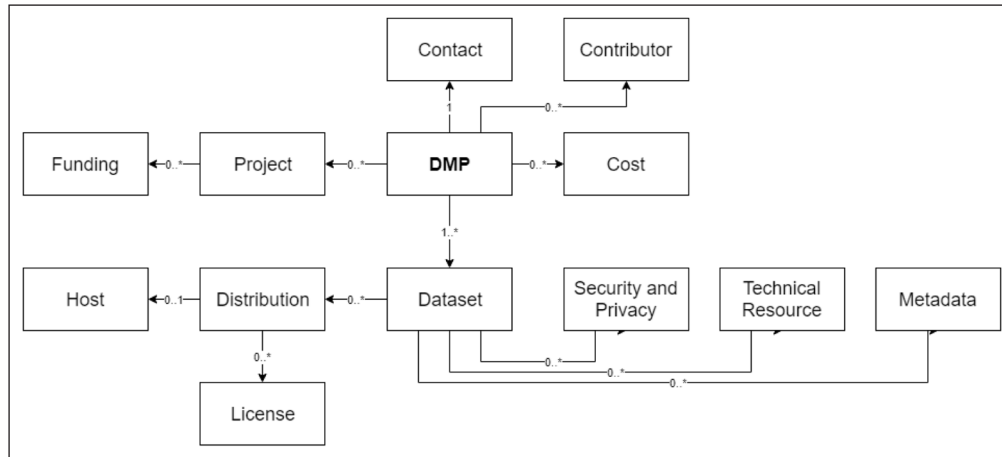

Figure 1 Overview of concepts and relations between them.

A minimal maDMP consists of only those fields that are defined as obligatory in the application profile. An example of a minimal maDMP compliant with the application profile is depicted in *Listing 1*. The listing shows that each maDMP must have a title. Each maDMP must also provide information on the contact person who can provide further information on data described by the maDMP. Each maDMP must indicate when it was created for the first time and when last modified. Thus, systems processing maDMPs can track their versions. Each maDMP must indicate the language in which it is written, so that its contents can be displayed properly

```
{
  "dmp": {
    "title": "Minimal DMP",
    "contact": {
      "contact_id": {
        "identifier": "http://orcid.org/0000-0002-4929-7875",
        "type": "orcid"
      },
      "mbox": "miksa@ifs.tuwien.ac.at",
      "name": "Tomasz Miksa"
    },
    "created": "2018-07-23T10:10:23.6",
    "dmp_id": {
      "identifier": "https://doi.org/10.0000/00.0.1234",
      "type": "doi"
    },
    "dataset": [
      {
        "dataset_id": {
          "identifier": "https://doi.org/10.0000/00.0.5678",
          "type": "doi"
        },
        "title": "Placeholder dataset",
        "personal_data": "unknown",
        "sensitive_data": "unknown"
      }
    ],
    "ethical_issues_exist": "unknown",
    "language": "eng",
    "modified": "2019-02-06T15:30:42.1"
  }
}
```

Listing 1 Example of a minimal maDMP in JSON.

(e.g. translated). Each maDMP must indicate whether any sensitive, personal or ethical issues exist. These fields follow a controlled vocabulary with three values to choose from: yes, no, unknown, so that no assumptions have to be made. Each maDMP must contain at least one dataset. A single dataset can represent "all data in a project" – like traditional DMPs often do. However, it is recommended (but optional) to use more datasets to provide more precise information.

5 DISCUSSION

This section provides discussion on the application profile. While the recommendation provides details on specific elements of the application profile, here we clarify selected aspects of using it and discuss design decisions taken.

5.1 CARDINALITY OF FIELDS

Only those fields for which the cardinality is set to "exactly one" or "one to many" must always be filled with information. Further fields defined in the application profile may be set if required (by business constraints), or when the information becomes available.

The application profile aims to be flexible and for this reason many fields are optional. In specific deployments requirements may be stricter, for example: DMP must contain information on a project number (funder requirement), while in the application profile specification this is optional.

All tools compliant to the application profile must expect to receive both obligatory and optional fields.

5.2 GRANULARITY OF DATASETS

There is no *single common* vocabulary that describes types of datasets. This shows that there is no consensus within the community on how to describe datasets. Sometimes, especially at the early stages of the research data lifecycle, it is necessary to refer to datasets on an aggregated or even abstract level, e.g. collection of satellite images. However, in other settings, one may assume that a dataset is equivalent to a file.

The application profile reuses the Dataset class as defined by the DCAT (Archer [2014]) which in turn is a sub-class of Dataset defined in DC Terms as: "Data encoded in a defined structure".[15] This definition is also very general. For this reason, we did not decide to constrain the definition of Dataset and its granularity depends on the specific context.

If a DMP contains only one Dataset (the most generic setting), it can denote that all data, for which the DMP is created, is considered jointly. For example, if a DMP is a short document created before a project begins and contains only an outline of planned actions.

If a DMP contains more than one Dataset, then each dataset can represent a logical group of data, e.g. raw data, software, etc. Thus, the application profile allows to express that different data is handled in a different way. For example, software is deposited in a source code repository under embargo, while data is instantly available in an institutional repository.

Distribution points to a specific instance of a dataset. Hence, distribution contains information like format and size of files.

A dataset can have several distributions. For example, an image can be both available as PNG and TIFF. Furthermore, a dataset can have many distributions to indicate where the data is kept temporarily, for example during a project, and where the data is going to be published/archived at the end of a project.

5.3 VERSIONING

Each DMP has a creation and a modification timestamp. The modification timestamp indicates the last modification of the DMP. Having two DMPs with different modification timestamps, one can identify which is newer by comparing timestamps. The same creation timestamp and DMP identifier indicate that we consider different versions of the same DMP.

---

15   *https://dublincore.org/specifications/dublin-core/dcmi-terms/#Dataset*.

The application profile itself does not have any mechanisms to model different versions of data – if information is overwritten, then previous information is not kept in the model. Systems processing DMPs must have suitable versioning mechanisms, if needed. For example, each update to a DMP can be committed to a database. Thus, the database engine allows to retrieve different versions of a DMP over time, while the DMP itself contains the modification timestamp allowing to identify/distinguish/refer to a specific DMP versions. The modification timestamp must be set by a tool that modified the DMP.

5.4 EMBARGO

An embargo on data sharing means that data will be made available using a license, but not immediately after deposition of data in a repository. As long as no license applies, the data is considered closed.

For each distribution, one can assign a license. If the license is assigned, then it means that a distribution at some point will become available. Start date set for the license indicates from when on it becomes active – in other words, when the distribution becomes available under this license.

We used the same mechanisms as defined by the National Information Standards Organisation.[16]

5.5 STATE OF THE DMP

We use dates to indicate planned actions. DMP has a modification timestamp, and Dataset contains issue date. Together these indicate whether the actions are planned or already performed.

- if the issue date is set in the future (compared to DMP modification date), then the actions are planned,
- if the issue date is set in the past (compared to DMP modification date), the actions were performed.

This approach is similar to the way embargo is modelled. We also wanted to avoid labels that can have different interpretations depending on the context. If we used a tag such as *approved* for a DMP, then someone could assume that the DMP was approved by the funder, whereas it was actually approved internally by the research support, before it was sent to the funder. Hence, we avoided modelling different states a DMP or a Dataset can be in, but focused on collecting facts, such as dates.

5.6 SERIALISATIONS

All the examples provided are in JSON, because of its readability and popularity, i.e. many tools implementing the application profile use the JSON serialisation. The application profile can be serialised to any other representation, e.g. XML, OWL, JSON-LD, etc. if needed. There is an ongoing work on the ontological representation of the application profile.

6 ADOPTIONS

This section describes existing adoptions of the application profile. Its goal is to provide concrete examples and further pointers to information on how the application profile can be used. The list of adoptions below is not final and further applications of the application profile are pending.

Furthermore, Table 2 (see Appendix A) presents an overview of systems adopting the recommendation. It provides a short description of each system, lists stakeholders interacting with it, outlines key benefits and explains in what setting it is used.

6.1 HAPLO

Haplo is based in the UK, and supplies research information management software for Higher Education. A suite of products cover the full research lifecycle, including Current Research

---

16 *https://www.niso-sts.org/TagLibrary/niso-sts-TL-1-0-html/element/ali-license_ref.html*.

Information System (CRIS) and repository functionality.[17] These products coexist within a single Haplo application, providing an integrated datastore for research activity and management at an institution. Haplo Repository[18] and the maDMP implementation[19] are open source.

Implementing a machine-actionable format enables Haplo to embed data management within the existing institutional research administration processes, sharing data between areas of the application. For example, an ethical approval process will update the project's DMP if the submission indicates sensitive data will be produced. The DMP is then queried when depositing a dataset to the repository, and will notify repository staff if the DMP indicates the dataset needs to be restricted (Renner [2020]) (illustrated in *Figure 2*). This implementation of maDMPs is in use at London South Bank University,[20] as is in the process of being deployed to further institutions at the time of writing.

Figure 2 Notification to library staff identifying a potential security issue when depositing a dataset to the repository.

The application profile is flexible, expressive, and general enough to be the native data structure for maDMPs within Haplo applications. This implementation of the application profile is in use at within a fully featured CRIS has demonstrated that the maDMP, structured correctly, can enable data management within the full rage of administrative processes that take place over the lifecycle of a research project. It enables the system to automatically apply policy, detect potential errors, and provide useful and meaningful reports on the data management activities at an institution.

6.2 OPEN RESEARCH PUBLISHING PLATFORMS

F1000 Research[21] is an open access publisher providing open research platforms to several funding partners including Wellcome, the Gates Foundation, and the Health Research Board of Ireland. These platforms use a unique open post-publication peer review model and are supported by an open and FAIR data sharing policy.[22] As part of F1000 Research's commitment to open research, there is a pilot project in place to extend this service offering to include the publication of DMPs. Metadata is a critical component of this project and especially in this context, it is key to openly sharing information across technologies, disciplines, and stakeholders. The end goal is to establish DMPs as project 'hubs' linking out to related research articles, datasets, etc. In this way, published DMPs will provide linkages which make connecting and tracking the research landscape easier – particularly for funders and institutions.

By adopting this application profile, F1000 Research plans to provide basic interoperability between systems producing or consuming machine-actionable DMPs while reducing redundancy of effort. A key advantage from F1000 Research's standpoint is that this application profile covers a broad range of use cases. Furthermore, it does not enforce any funder or institutional specific requirements. It is also extremely important that the application profile represents information over the whole DMP lifecycle, given that each published DMP will be versionable.

---

17   *https://haplo.com*.

18   *https://github.com/haplo-org/haplo-repository*.

19   *https://github.com/haplo-org/haplo-research-manager/tree/master/hres_data_management_plans*.

20   *https://www.haplo.com/news/haplo-repository-dmp*.

21   *http://f1000research.com*.

22   *https://f1000research.com/for-authors/data-guidelines*.

6.3 DMPTOOL

The cornerstone of support for DMPs in the U.S. is the DMPTool, developed in 2011 by the California Digital Library (CDL) and founding collaborators.[23] The Digital Curation Center, based in the UK, and CDL have a formal partnership to co-develop and maintain a single, open-source platform, for providing DMP guidance. This shared software, DMPRoadmap, underpins both the DMPTool and the DCC service, DMPonline.[24]

In 2017 CDL was awarded a National Science Foundation EAGER grant to support the creation of machine-actionable DMPs.[25] As part of this grant, the DMPTool, together with the DMPRoadmap team, implemented the application profile within the shared code base and developed an application profile compliant API. Additionally, in partnership with DataCite, DMPTool developers have built a workflow for generating DOIs for DMPs and utilising the Event Data service from DataCite to record when assertions have been made on the DOI.

Ongoing pilot projects with domain-specific and institutional stakeholders are testing integration between different services and systems utilising maDMPs. The largest of these pilot projects is the FAIR Island project which is a collaboration between the University of California Gump Field Station, located on Moorea in French Polynesia, and Tetiaroa Society, which operates a newly established field station located on the atoll of Tetiaroa. By implementing mandatory registration requirements including extensive use of controlled vocabularies, PIDs, and other identifiers, maDMPs in this environment will be utilised as key documents for tracking provenance, attribution, compliance, deposit, and publication of all research data collected on the island. The FAIR Island Project offers a real-world example to prove the capabilities of machine-actionable DMPs and to analyse the downstream effects of these policies in the resulting release of data.[26]

6.4 DMPONLINE

The Digital Curation Centre launched DMPonline[27] in 2010 to support the UK Higher Education sector with the increasingly complex landscape of divergent funder policies for DMPs. Over the years the user base has expanded significantly, and the service now has a growing number of paying subscribers at institutions and funders in the UK, Netherlands, Sweden, Finland and Australia.

Together with partners at the CDL, the DCC is enhancing the open-source DMPRoadmap codebase to support machine-actionable DMPs. One of the first enhancements in this series of activities pre-dates the DMP Common Standards Working Group. In 2016, the DCC received an RDA Europe collaboration award to integrate the Metadata Standards Directory into DMPonline. By using external directories as a way to answer certain questions, we could ensure structured data was provided rather than free text. Work to integrate other registries of repositories and standards such as Re3data and FAIRsharing continues, and integrations with the Research Organisations Registry and FundRef are planned.

The DMPRoadmap data model instantly complied with the application profile for maDMPs at a minimal level (Rust [2019]) but as noted above, much work has been done to extend this, for example by adding additional metadata such as project start/end dates, utilising a range of persistent identifiers and developing an application profile compliant API. The existing DMP themes[28] which are used as tags on questions and guidance in the system also provide an approximation to other key elements of the application profile such as *ethical issues*, *metadata*, *costs* and *preservation statement*. Mapping between these would permit existing data from the existing 80,000 DMPs to be converted to a more comprehensive maDMP for reuse in other

---

23   DMPTool founding collaborators include: Digital Curation Centre (DCC), DataONE, Smithsonian Institution, University of California San Diego, University of California Los Angeles, University of Virginia, and University of Illinois at Urbana-Champaign.

24   *https://github.com/DMPRoadmap*.

25   *https://www.nsf.gov/awardsearch/showAward?AWDID=1745675HistoricalAwards=false*.

26   For more information on the FAIR Island Project visit *fairisland.org*.

27   *https://dmponline.dcc.ac.uk*.

28   *https://github.com/DMPRoadmap/roadmap/wiki/Themes*.

systems. In a recent RDA Europe hackathon, developers from DMPTool and DMPonline trialled the new API and supported teams from Haplo, the Data Stewardship Wizard, OpenDMP and others to both export and import maDMPs from DMPRoadmap.

6.5 DMP OPIDoR

DMP OPIDoR is a DMP tool, based on DMPRoadmap code. It is made available to the French scientific community and has been adapted to meet its needs.

In France, commitment to open science is widely embraced by all stakeholders who are involved at different stages of the data lifecycle. They are willing to provide their expertise and interconnect their infrastructures to DMPs so as to enable a seamless data management.

In order to ease information exchange between systems, a new flexible object data model has been developed based upon the DMP templates currently published in DMP OPIDoR and use cases submitted by the different actors. Its implementation into DMP OPIDoR is underway. As a result, more structured and standardised information will be collected, and the use of PIDs and controlled vocabularies will also be facilitated. Furthermore, the flexibility and extensibility of this model will allow it to be adapted to the specifics of various service providers and scientific communities.

These developments will be first tested as part of a project, in partnership with the French Bioinformatics Institute and funded by the French Research National Agency, whose main objective is to enable automated service provisioning by a data processing facility.

The implemented data model is compliant with the application profile described in this paper. DMP OPIDoR will provide an import/export feature to allow interoperability with other DMP tools or publishing platforms.

6.6 DATA STEWARDSHIP WIZARD

Data Stewardship Wizard (DSW, Pergl et al. [2019])[29] as a data management planning tool has a significantly different approach compared to others: it is not meant primarily to satisfy demands by funders and institutions, but to help researchers plan data management that works best for their project. Rather than using question templates directly, DSW makes use of customizable knowledge models that describe the hierarchical structure and content of questionnaires. DSW comes with a *root* knowledge model that avoids free-text answers, and contains guidance towards making data FAIR. A questionnaire can be turned into a document using Jinja2[30] templates that can either simply print-out the questions and answers (e.g. the generic *default template*), or synthesise text from the answers (e.g. following the Science Europe template), or transform answers to any textual format following a required schema (e.g. RDF, JSON, or OPML).

Machine-actionable DMPs are the key to interoperability between DMP tools and potentially will enable use of structured information from DMPs by different stakeholders such as project management offices, institutional IT support, and funders. To support this, we implemented export and import functionality in DSW for maDMPs compliant with the application profile. We achieved export functionality using Jinja2 templates for creating documents. The main part of this task was then to find a mapping between the application profile and the *root* knowledge model of DSW. Many questions were already present and directly mappable in the *root* knowledge model, e.g. several questions regarding ethical issues, some other questions had to be modified or moved in the structure, e.g. so that a maDMP can describe several projects instead of just one according to the application profile. Also, some new questions were added, e.g. to support describing the funding of each project where we also added an integration with the funders registry of CrossRef.[31] Those changes together were released as a new version of the *root* knowledge model, and anyone with an existing questionnaire can migrate their answers. The newly added maDMP export template queries the relevant answers in a questionnaire using their UUIDs and transforms them into a JSON object according to the DCS JSON schema

---

29   *https://ds-wizard.org*.

30   *https://jinja.palletsprojects.com/en/2.11.x/templates/*.

31   *https://www.crossref.org/services/funder-registry/*.

and dumps it as output. In addition to a JSON export, we also created a template for RDF using the ontology representation of the application profile. It uses the mentioned JSON object to compose an RDF Turtle file that can be then internally converted to other RDF formats such as RDF/XML, Trig, JSON-LD, or N3.

Implementing the import of maDMPs was more complex as there was no concept of importing questionnaires in our system beforehand. Nevertheless, we were able to create a prototype import functionality on the frontend part of DSW. After loading a valid JSON file, it decodes the maDMP into the internal structure and then, using the same mapping used for the export, transforms the data into answers for a new questionnaire. The user can preview the parsed data and then create such a pre-filled questionnaire. Within this prototype the structure of maDMPs as well as the mapping was hard-coded; we will use this for further analysis and generalisation into a future flexible answer-import feature. By adopting maDMPs (as shown in *Figure 3*), we managed to interchange maDMPs with other DMP tools, such as DMPtool, Argos, and easyDMP. Moreover, using the *document submission* feature of DSW, it is possible for a user to send a DMP from DSW to DMPTool via its API by clicking a button in DSW.
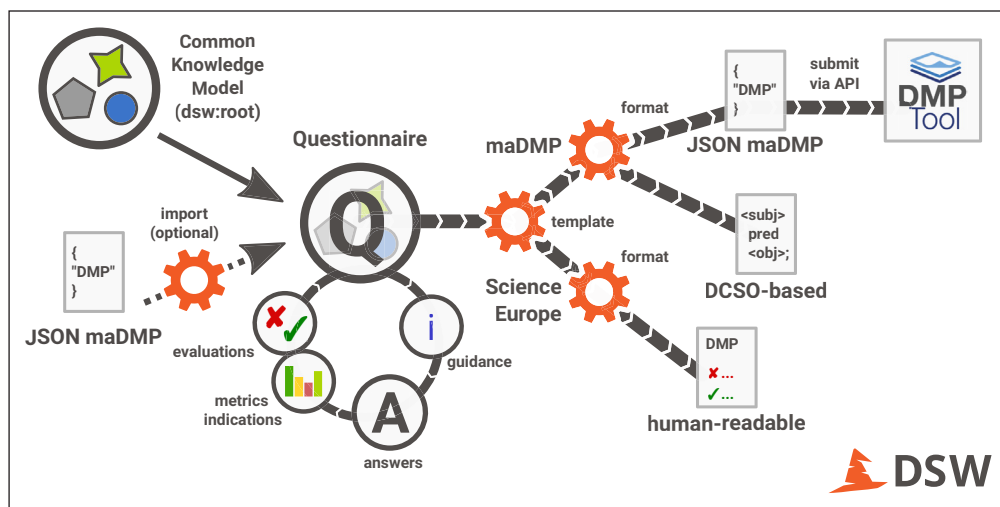

Figure 3 Diagram of how DSW is using maDMPs.

6.7 NSD DMP

NSD – Norwegian Centre for Research Data – is a national archive and research data centre. Its mission is to ensure free and open access to research data, and improve the basis for empirical research through a broad range of data and support services. By means of automation and system integration, the NSD DMP can assist researchers and institutions in their efforts to share their data.

NSD DMP adopts parts of the application profile. The NSD specifically looked at new ways of distinguishing between project information and dataset information when creating a DMP. We also considered ways of integrating security and privacy issues with information on relevant data hosting options. We found much valuable input and inspiration in the set of semi-automated workflows that were designed in the RDA DMP Common Standards Working Group. These processes show how data management planning can be supported by means of automation and system integration in an institutional context.

An example of machine-actionability in the NSD DMP is the classification module, which provides institutional-specific policy recommendations for collecting, storing and archiving data based on the classification of data into either open (green), restricted/internal (yellow), confidential (red), or strictly confidential (black) categories. Another example is the archive guide, which helps users identify national archives and repositories that are relevant for their data. The archive guide uses APIs from *re3data.org*, and provides a list of suggested archives and repositories for data packages based on various criteria pre-selected by the user.

In addition the NSD DMP tool is integrated with a Data Policy Manager, which allows institutions to design interactive and machine-actionable policies that can be linked to internal and external systems as needed. The tool enables institutions to define their own policies for a

wide variety of general and institution-specific storage services, data transfer applications and data collection tools.

The proof of concept tools and BPMN processes that were developed together with the application profile helped us to better imagine the full scope of a system and provided us with valuable suggestions as on how to develop our own DMP tool with machine-actionable elements.

6.8 ARGOS – OPENDMP

Argos is a service offered by OpenAIRE[32] utilising the OpenDMP platform[33] co-designed by OpenAIRE and EUDAT initiatives, in order to support the management and distribution of machine actionable DMPs.

The platform builds on a few core concepts, which are as follows:

- Dataset description: structured and unstructured (i.e. textual) information about the dataset, shaped by a dataset Profile
- Data Management Plan: an aggregation (1.n) of dataset descriptions into a set managed as a whole in a given context
- Dataset Profile: a definition, containing data elements and behavioural rules, that dictates how a dataset may be described by a user

Around those baseline concepts, numerous other "native" (i.e. a-priori defined in code) and "soft" (i.e. defined by configuration) elements are utilised to complete the data model of the system. Examples of those elements are projects, funders, researchers, organisations, repositories etc.

OpenDMP supports maDMP JSON import and export under a few assumptions. Its model maps natively to a substantial segment of the maDMP entities, attributes and relations. There is the DMP container which aggregates a number of datasets, contributors, contacts and language information, while DMP timeline attributes are captured via a combination of DMP versions and timestamps. The project, grant and funder stack of the application profile is supported under a slightly different arrangement which fits well when exporting maDMPs, yet imports may lead to a user prompt for further choices. Two deviations are that: (i) OpenDMP supports the actual funding, the cost and ethical issues at the level of a dataset instead of the entire DMP, and as such the maDMP elements may be exported only in the case of DMPs with a single dataset. (ii) Only one distribution is supported per dataset, which works well during export but imports may lead to user prompts for further choices. Beyond those differences all attributes of maDMP dataset model are covered by OpenDMP and may both exported. Through its customisable templates, OpenDMP may support concepts not directly mapped to maDMPs. In order to enable maDMP JSON format as its fully re-importable export form, it utilises a reserved area of the file for maintaining the extra information. The other way round, OpenDMP imports maDMP files under certain assumptions (as presented above) and tries to reserve additional data and structure that may not fit its model in reserved areas to avoid discarding the extra information. This extra info remains accessible through its REST API. Currently work is underway to fine-tune the mapping of particular attributes to maDMP concepts and extending the user actions supported during imports.
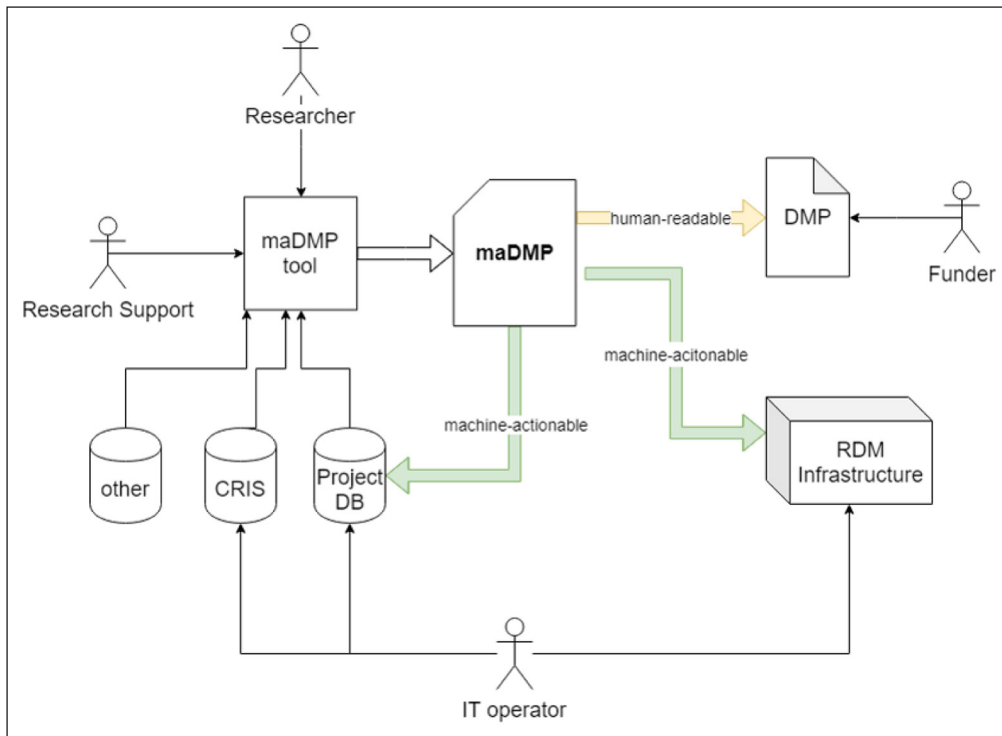
6.9 RESEARCH DATA INFRASTRUCTURE AT TU WIEN

TU Wien is the largest technical university in Austria. Within the FAIR Data Austria project[34] TU Wien is developing an integrated infrastructure for research data management that includes data and code repositories, as well as tools for machine-actionable DMPs. The goal is to integrate existing and new systems to ensure exchange of information and improve management of scientific data (see *Figure 4*).

---

32   *https://argos.openaire.eu/*.

33   *https://gitlab.eudat.eu/dmp/OpenAIRE-EUDAT-DMP-service-pilot*.

34   *https://forschungsdaten.at*.

Figure 4 maDMPs used to exchange machine-actionable DMPs among university services.

Machine-actionable DMPs are one of the main enablers of the project. They act as an inventory integrating information from various systems. For example, maDMPs provide a link between research groups, projects and data produced by them. The maDMPs help in realising 'ask once only' principle that aims at reducing the workload imposed on researchers and support services. For example, information on existing ethical issues is requested when a new project is registered. This information is automatically transferred to a DMP. TU Wien is currently developing a tool for DMPs that will be shared with project partners: TU Graz and University of Vienna. The application profile and prototypes developed in course of the research streamlined the architecture of the developed data infrastructure and are being implemented in live systems.

6.10 EASYDMP

The easyDMP tool[35] for creating DMPs was created in Norway in 2016 by UNINETT Sigma2 to provide a simple Python tool for creating DMPs that integrates with the services for provisioning storage on the Norwegian Infrastructre for Research Data (Nird) operated by Sigma2. The tool supports a variety of templates (Horizon 2020, Science Europe and local templates) that are defined through an administration interface. The level of detail is defined by the template designer and can have branches and make use of controlled vocabularies. Plans are stored in an SQL database and the information can be accessed through a read-only API.

The goal of easyDMP has been make the plan machine actionable enabling storage services to reserve the storage space requested in the plan. The development of the application profile by the RDA working group has been fortuitous as it has provided a schema based on feedback from a large number of communities interested in providing maDMPs.

By aligning easyDMP with the application profile we can ensure our tool interoperates with other tools making the plans independent of the tool implementation. We are also working to extend the schema such that our storage services can consume the plans and reserve the required storage space. Providing a different DMP tool implements our extension our storage services would be able to consume plans created in that tool.

7 CONCLUSION

We can reduce the effort imposed on researchers and on other stakeholders involved in research data management, when information contained in data management plans can be accessed,

---

interpreted, exchanged and acted upon by machines. Automation and machine-actionability can lead to improved quality of information provided and higher reuse of information resulting in better return on invest made into data management services.

In this paper we reported on a community effort to define the application profile for machine-actionable data management plans that allows expressing information from traditional data management plans in a machine-actionable way. We described the methodology and research conducted to define the application profile. We also discussed design decisions made during its development and presented systems adopting it that include major DMP tool providers, as well as repositories, publishers and universities. The application profile is an official output of the RDA DMP Common Standards Working Group that has gathered more than 200 members from around the globe.

The next steps will focus on further adoptions of the application profile and development of novel services utilising the full expressiveness of the application profile, e.g. repository recommendation services and other services acting on behalf of stakeholders. We also plan to release new serialisations of the application profile, such as the OWL ontology to fully utilise the benefits of machine-actionability and the Semantic Web.

The RDA DMP Common Standards Working Group will continue to maintain the application profile and will incorporate feedback received from the adopters. Any updates to the specification of the application profile will remain domain and tool independent, so that the application profile remains generally applicable.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Appendix A.** Overview of adoptions. DOI: *https://doi.org/10.5334/dsj-2021-032.s1*

ACKNOWLEDGEMENT

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Tomasz Miksa *orcid.org/0000-0002-4929-7875*
TU Wien & SBA Research, Austria

Paul Walk *orcid.org/0000-0003-1541-5631*
Antleaf Ltd., UK

Peter Neish *orcid.org/0000-0002-7171-2334*
The University of Melbourne, Australia

Simon Oblasser *orcid.org/0000-0002-3611-3483*
TU Wien, Austria

Hollydawn Murray *orcid.org/0000-0002-8243-2493*
F1000 Research, UK

Tom Renner *orcid.org/0000-0001-9500-4407*
Haplo, UK

Marie-Christine Jacquemot-Perbal *orcid.org/0000-0002-6316-1472*
INIST-CNRS, France

João Cardoso *orcid.org/0000-0003-0057-8788*
INESC-ID & Instituto Superior Técnico, Portugal

Trond Kvamme *orcid.org/0000-0002-8657-7833*
NSD – Norwegian Centre for Research Data, Norway

Maria Praetzellis *orcid.org/0000-0001-5047-3090*
University of California, California Digital Library, United States

Marek Suchánek ⓘ *orcid.org/0000-0001-7525-9218*
Czech Technical University in Prague, CZ

Rob Hooft ⓘ *orcid.org/0000-0001-6825-9439*
Dutch Techcentre for Life Sciences, Netherlands

Benjamin Faure ⓘ *orcid.org/0000-0003-4521-9252*
INIST-CNRS, France

Hanne Moa ⓘ *orcid.org/0000-0002-6635-0033*
Uninett, Norway

Adil Hasan ⓘ *orcid.org/0000-0003-2373-7112*
Uninett Sigma2, Norway

Sarah Jones ⓘ *orcid.org/0000-0002-5094-7126*
Digital Curation Centre, UK

REFERENCES

Aigner, MB. 2020. Proof of Concept tool: Making maDMPs human readable. DOI: *https://doi.org/10.5281/zenodo.3727714*

Alkhatib, H and **Rivera, C.** 2020. Proof of Concept tool: Making maDMPs human readable. DOI: *https://doi.org/10.5281/zenodo.3727724*

Archer, P. 2014. Data catalog vocabulary (dcat) (w3c recommendation), Online. URL: *https://www.w3.org/TR/vocab-dcat/*.

Bakos, A, Miksa, T and **Rauber, A.** 2018. Research data preservation using process engines and machine-actionable data management plans. In: '*TPDL', Vol. 11057 of Lecture Notes in Computer Science*, 69–80. Springer. DOI: *https://doi.org/10.1007/978-3-030-00066-0_6*

Breitenfellner, H. 2020. Proof of Concept tool: RDM Organizer to maDMP export. DOI: *https://doi.org/10.5281/zenodo.3727753*

Dalpiaz, F and **Brinkkemper, S.** 2018. Agile requirements engineering with user stories. In: '*2018 IEEE 26th International Requirements Engineering Conference (RE)'*, 506–507. DOI: *https://doi.org/10.1109/RE.2018.00075*

DCC. 2013. Checklist for a Data Management Plan. v.4.0. Edinburgh: Digital Curation Centre. *http://www.dcc.ac.uk/resources/data-management-plans*. Online; accessed 29 March 2018.

Doorn, P. 2018. *Science Europe Practical Guide to the International Alignment of Research Data Management.*

Engelhardt, C, Enke, H, Klar, J, Ludwig, J and **Neuroth, H.** 2017. Research data management organiser. In: *Proceedings of the 14th International Conference on Digital Preservation*. iPRES 2017, Kyoto, Japan, September 25–29, 2017.

Hevner, AR, March, ST, Park, J and **Ram, S.** 2004. Design science in information systems research. *MIS Q*, 28(1): 75–105. DOI: *https://doi.org/10.2307/25148625*

Hido1994 and **Alhirthani, A.** 2020. Dataverse and maDMPs integration PoC. DOI: *https://doi.org/10.5281/zenodo.3727707*

Inschlag, L and **Drechsel, M.** 2020. Proof of Concept tool: DMP Roadmap to maDMP export. DOI: *https://doi.org/10.5281/zenodo.3727757*

Jones, S, Pergl, R, Hooft, R, Miksa, T, Samors, R, Ungvari, J, Davis, RI and **Lee, T.** 2020. Data management planning: How requirements and solutions are beginning to converge. *Data Intelligence*, 2(1–2): 208–219. DOI: *https://doi.org/10.1162/dint_a_00043*

Leidinger, M. 2020. Proof of Concept tool: Making maDMPs human readable. DOI: *https://doi.org/10.5281/zenodo.3727720*

Michener, WK. 2015. Ten simple rules for creating a good data management plan. *PLoS Comput Biol*, 11. DOI: *https://doi.org/10.1371/journal.pcbi.1004525*

Miksa, T, Cardoso, J and **Borbinha, JL.** 2018. Framing the scope of the common data model for machine-actionable data management plans. In: Abe, N, Liu, H, Pu, C, Hu, X, Ahmed, NK, Qiao, M, Song, Y, Kossmann, D, Liu, B, Lee, K, Tang, J, He, J and Saltz, JS (eds.), *IEEE International Conference on Big Data*. Big Data 2018. Seattle, WA, USA, December 10–13, 2018. IEEE, 2733–2742. DOI: *https://doi.org/10.1109/BigData.2018.8622618*

Miksa, T, Neish, P, Walk, P and **Rauber, A.** 2018. Defining requirements for machine-actionable data management plans. In: McGovern, N and Whiteside, A (eds.), *Proceedings of the 15th International Conference on Digital Preservation*. iPRES 2018. Boston, MA, USA, September 24–28, 2018. URL: *https://hdl.handle.net/11353/10.923628*.

Miksa, T, Rauber, A, Ganguly, R and **Budroni, P.** 2017. Information integration for machine actionable data management plans. *IJDC*, 12(1): 22–35. URL: *http://www.ijdc.net/index.php/ijdc/article/view/12.1.22*.

Miksa, T, Simms, S, Mietchen, D and **Jones, S.** 2019. Ten principles for machine-actionable data management plans. *PLOS Computational Biology*, 15(3): 1–15. DOI: *https://doi.org/10.2218/ijdc.v12i1.529*

Miksa, T, Walk, P and **Neish, P.** 2019. RDA DMP Common Standard for Machine-actionable Data Management Plans. DOI: *https://doi.org/10.1371/journal.pcbi.1006750*

Oblasser, S. 2019. Machine-actionable dmp application (dmap). The slides are based on a live demo given at RDA 14th Plenary taking place in Helsinki between 22–25 October 2019. Further session material can be found here: *https://www.rd-alliance.org/p14-helsinki-slides*. DOI: *https://doi.org/10.5281/zenodo.3522247*

Oblasser, S. 2020. oblassers/dmap-mockups: Dmap mockups. DOI: *https://doi.org/10.5281/zenodo.3630375*

Oblasser, S and **Miksa, T.** 2019. BPMN Processes for machine-actionable DMPs. DOI: *https://doi.org/10.5281/zenodo.2607556*

Oblasser, S, Miksa, T and **Kitamoto, A.** 2020. Finding a repository with the help of machine-actionable DMPs: opportunities and challenges. DOI: *https://doi.org/10.5281/zenodo.3701564*

Pergl, R, Hooft, R, Suchánek, M, Knaisl, V and **Slifka, J.** 2019. Data Stewardship Wizard: A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning. *Data Science Journal*, 18(1). DOI: *https://doi.org/10.5334/dsj-2019-059*

Pichler, M. 2019. martinpichler/dmp-roadmap-parser: Release version 0.1. DOI: *https://doi.org/10.5281/zenodo.3250398*

Renner, T. 2020. Turbo-charging data management plans. DOI: *https://doi.org/10.5281/zenodo.3673058*

Rust, S. 2019. Implementing maDMPs in DMPonline, RDA Webinar. URL: *https://www.rd-alliance.org/rda-output-adoption-webinar-series-data-management-plans*.

Simms, S, Jones, S, Mietchen, D and **Miksa, T.** 2017. Machine-actionable data management plans (madmps). *Research Ideas and Outcomes*, 3: e13086. DOI: *https://doi.org/10.3897/rio.3.e13086*

Simms, SR, Jones, S, Miksa, T, Mietchen, D, Simons, N and **Unsworth, K.** 2018. A landscape survey of activedmps. *IJDC*, 13(1): 204–214. DOI: *https://doi.org/10.2218/ijdc.v13i1.629*

Smale, N, Unsworth, K, Denyer, G and **Barr, D.** 2018. The history, advocacy and efficacy of data management plans. *bioRxiv*. URL: *https://www.biorxiv.org/content/early/2018/10/17/443499*. DOI: *https://doi.org/10.1101/443499*

Tsepelakis, S. 2020. SotosTsepe/invenio-madmp: First pre-alpha release of invenio-madmp. DOI: *https://doi.org/10.5281/zenodo.3733268*