Framing the scope of the common data model for machine-actionable Data Management Plans

Tomasz Miksa
SBA Research & TU Wien
Vienna, Austria
tmiksa@sba-research.org

João Cardoso
INESC-ID
& Instituto Superior Técnico
Lisbon, Portugal
joao.m.f.cardoso@tecnico.ulisboa.pt

José Borbinha
INESC-ID
& Instituto Superior Técnico
Lisbon, Portugal
jlb@tecnico.ulisboa.pt

***Abstract*— Currently, research requires processing data at a large scale. Data is not anymore a collection of static documents, but often a continuous stream of information flowing into information systems. Researchers need to manage their data efficiently not only to keep it safe, but also to ensure that it can be later correctly interpreted and reused. Existing solutions are not sufficient. Traditional Data Management Plans are manually created text documents that describe how research data will be handled. Yet, researchers must implement all actions by themselves. Machine-actionable Data Management Plans are a new approach that allows systems to act on behalf of researchers and other stakeholders involved in data management, to help them manage data in an efficient and scalable way. This paper summarises the results of work performed by the Research Data Alliance working group on Data Management Plan Common Standards to realise this vision. The paper describes results of consultations and proof of concept tools that help in: identifying needs for information of stakeholders involved in data management; defining the scope of the common data model for Machine-actionable Data Management Plans to allow for exchange of information between systems; identifying necessary services and components of infrastructure that support automation of data management tasks.**

***Index Terms*— ata Management Plan, Machine-Actionable Data Management Plan, Workflowsata Management Plan, Machine-Actionable Data Management Plan, WorkflowsD**

I. Introduction

With advances in technology, scientific research requires data processing in an increasingly larger scale. Data is no longer a collection of static documents, but often a continuous stream of information flowing into a repository [1], for example, satellite images or sensor data captured periodically. This new paradigm of research is often described as e-Science [2].

Researchers need to plan and manage their data efficiently, not only to keep it safe, but also to ensure that it can be later correctly interpreted and reused. This is especially important in the context of open data [3] and FAIR principles [4], [5].

One of the tools introduced to solve research Data Management (RDM) [6] challenges is the Data Management Plan (DMP) [7]. The overall objective of a DMP is to document, in a project, the techniques, methods and policies on how data is to be created, documented, accessed, preserved and disseminated. Various funding bodies, such as for example The National Science Foundation (NSF) or the European Commission (EC), already require that any funding application be accompanied by a DMP.

However, proper research data management, especially in view of big data and complex processing pipelines, is a complex task that requires cooperation of several stakeholders: not only researchers, but also, infrastructure operators, repository managers, legal experts, and so on. Researchers simply do not have enough expertise, nor time to prepare a DMP and then to actually implement it.

For this reason, there is a need for a solution that supports researchers in planning and managing data in an automated and scalable way. Research Data Alliance (RDA)[1] working group on DMP Common Standards[2] works to implement machine-actionable DMPs (maDMPs) [8]. The larger goal is to improve the experience for all involved by exchanging information across research tools and systems and embedding DMPs in existing workflows. As a result, parts of the DMP can be automatically generated and shared with other collaborators or funders. To achieve this goal there is the need for: good understanding of research data workflows, RDM infrastructure, common data model for maDMPs.

This paper presents the results to date of the RDA DMP Common Standards working group on realising maDMPs. It describes consultations performed and proof of concept tools developed that help in:

1) **identifying stakeholders involved in data management and their requirements for information;**
2) **narrowing the scope of the common data model for maDMPs that acts as a standard for exchange of information between systems involved in data management;**
3) **identifying necessary services and components of infrastructure that support automation of data management tasks.**

The paper is organised as follows. Section II provides definitions of the concepts of RDM, DMP, maDMP, and the RDA DMP Common Standards working group. Section III describes the work towards the creation of a DMP common model. Particular focus is given to the description of the two user consultations that were made to gather requirements for the development of the data model. In section IV we describe three tools that were developed as proof of concept, to demonstrate how a common DMP model can be used to automate tasks. Conclusions and outlook appear in section V.

II. Fundamentals

A. Research Data Management

As researchers must cope with the management of mounting quantities of data, and abide by the FAIR principles [4] of having data be findable, accessible, interoperable and reusable.

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

Research Data Management (RDM) [6] has taken a central role in scientific research [9].

RDM, sometimes also referred to as Scientific Data Management (SDM), is one of the approaches available to researchers, to tackle the challenge of how to manage, preserve and publish their data in way that allows for it to be reproduced and reused A definition of RDM is that it "concerns the organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and allows for new and innovative research built on existing information" [10].

RDM is best perceived when illustrated through the data life cycle [6], as seen in Figure 1.

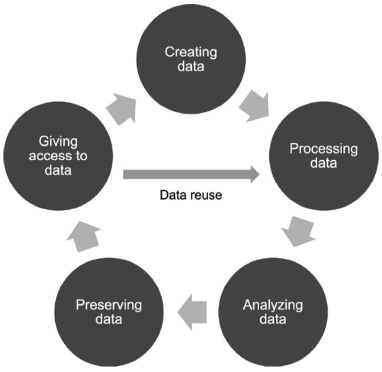


Fig. 1: Data life cycle [6]

The first three stages on life cycle focus on the creation or collection of raw data, that is then processed and analysed, so that any results can be made available through publishing. The latter two stages, deal with the preservation and access to the data. Thus allowing its results to be reproduced and reused by other researchers.

In order to cater for the needs of RDM, researchers need to ensure that the necessary scientific infrastructure [11] is in place to support RDM practices and policies. This scientific infrastructure does not refer solely to necessary physical facilities, resources, both human and material, and services [12] necessary to enforce RDM practices and policies, must also be considered.

B. Data Management Plan

The concept of Data Management Plan (DMP) was introduced to document and publish any RDM practices and policies that are applied to data throughout the duration of a research project. A DMP is a document detailing how data from a research project is to be managed throughout its life cycle. This implies describing the techniques, methods and policies on how data is to be created, documented, accessed, preserved and disseminated.

According to literature [7], there are 10 principles and practices that should be observed in the creation of a DMP. They are:

1) Determine the research sponsor requirements;
2) Identify the data to be collected;
3) Define how the data will be organised;
4) Explain how the data will be documented;
5) Describe how data quality will be assured;
6) Present a sound data storage and preservation strategy;
7) Define the project's data policies;
8) Describe how the data will be disseminated;
9) Assign roles and responsibilities;
10) Prepare a realistic budget.

C. Machine-Actionable Data Management Plan

Unfortunately the reality of DMP application does not match its conceptual vision. A DMP is meant to be created at the beginning of a research project. It should then be revisited and updated throughout the life cycle of the research data to witch it pertains. However, in practice DMP documents are created either at the onset or close of a project and are rarely revisited for updates.

In addition to this, the level of detail of a DMP varies according to the expertise, meticulousness of its creators or the funding agency template that is being provided. The lack of standardisation and consistency may lead to the DMP not having the necessary information, or offering broad or unclear descriptions of the RDM principles and practices that it is describing.

This results into a mostly static document. Overall these fact lead to the general perception of the DMP, as a static document and nothing more than bureaucratic hassle. Very far from being a relevant artefact for RDM.

The concept of maDMP [8] (sometimes referred as "active, "dynamic", or "machine-readable" DMP) was thus born from an initiative to tackle these limitations and introduce dynamic and machine-readable features to a DMP.

As is the case for the DMP, there is also a proposal for a set of 10 principles and practices that are specific to the maDMP [13]. These principles and practices were conceived to both aid in the application of the maDMP concept, as well as, to realise its benefits. They are:

1) Integrate DMPs with the workflows of all stakeholders in the research ecosystem;
2) Allow automated systems to act on behalf of stakeholders;
3) Make policies (also) for machines, not just for people;
4) Describe, for both machines and humans, the components of the data management ecosystem;
5) Use PIDs and controlled vocabularies;
6) Follow a common data model for maDMPs;
7) Make maDMPs available for human and machine consumption;
8) Support data management evaluation and monitoring;
9) Make DMPs updatable, living, versioned documents;
10) Make DMPs publicly available.

All these principles and practices would add value to the already existing DMP. However, due to the very nature of the machine-readable representation that is key to the maDMP concept, services can be developed to take advantage of the information described in the maDMP. This could enable parts of the DMP to be automatically generated and shared. Thus tackling the current limitation of the DMP when it comes to its lack of standardisation and consistency. Moreover, the possibility to automate DMP generation can also be extended to automate DMP updates. With automated updates, the DMP could truly become a an valuable artefact for RDM. Ultimately it would allow for its information to be discovered, shared and reused by other services or researchers.

D. RDA DMP Common Standards Working Group

During the 9th RDA plenary meeting in Barcelona, the need for a dedicated working group that would focus on standardising the knowledge contained within a DMP was identified. The DMP Common Standards working group [14] was created soon after, with an overall objective of establishing a common data model that defines a core set of elements for a DMP.

The proposed data model is to have a modular design, so as to allow for customisation and extensions by existing standards and vocabularies, in accordance to the best practices of different research communities. The information model is to use semantic

technologies, that have already been established as a viable solution in data management and preservation domains [15].

In order to achieve this objective, the DMP Common Standards Working group had first to define the necessary requirements for a maDMP. Through discussions with interested stakeholders it was concluded that there was neither a clear and common definition of maDMP, nor a consensus regarding the information it should contain.

III. Open Consultations

This section describes two consultations performed by the RDA DMP Common working group that aimed at clarifying the scope of the common data model for machine-actionable DMPs. The first consultation (see Section III-A) went broad and focused on identifying who and when needs what information, and who can actually provide this information. The second consultation (see Section III-B) went deep and allowed breaking down identified requirements into more fine granular fields that must become either parts of the core model or one of its extensions.

A. First Consultation

The first consultation aimed at answering following questions:

- Who are stakeholders at each lifecycle stage?
- How available information changes over the lifetime of a DMP?
- How need for information changes over the lifetime of a DMP?

The user consultation took place between the 9th of October 2017 and the 30th of November of 2017, and it is described in depth in [16]. Participants were gathered through direct invitation during RDA events, workshops aimed at non RDA stakeholders and twitter.

The user consultation was realised through the application of the user stories approach [17]. This approach is applied in software development to gather feature descriptions, expressed in natural language, offering a perspective of end users. User stories typically follow the following template: As a <stakeholder>, I want <goal> so that <reason>. Once written, user stories can then be classified into functional or non-functional requirements by system/software architects.

There were two main criteria that were considered when selecting an approach for this user consultation. The large number of community members that were to be consulted, and the fact that the selected approach should not require expert training. The user stories approach fulfils both criteria. It can be applied to a wide range of participants, and does not require any training other than the participant's own views and experience on the consultation topic.

The tool chosen to conduct the user consultation was GitHub[3]. Through its issue mechanism, the community was able to contribute with suggestions and resources. Participants were allowed to both contribute with their user stories, but also evaluate and comment on the stories other participants submitted. An example of this process can be seen in Figure 2.

After the user consultation was closed, all submitted user stories were then organised according to three categories: *Accepted*, when their content could directly be related with the building of the data model. *Out of scope*, when their content could not be proved to be within the scope of the working group. And finally, *ignored/rejected* when the submitted user story had no relation with the context under analysis. In total 180 user stories were collected, of which 77 were marked as accepted, 22 as being out of scope and 9 as rejected.

Additionally the user stories were also classified, using coloured labels, in accordance to their stakeholder, project phase and subject of information conveyed in a DMP. Examples of this classification can be seen in the various user stories displayed in Figure 2.

The combination of the organisation and classification processes allowed for requirements[4] to be drawn from the user stories. However, through analysis of the created requirements it was noted that distinct contexts were producing similar requirements. As a result the full set of requirements were organised into five categories:

- Administration roles and responsibilities[5];
- Data[6]
- Infrastructure[7];
- Security, privacy and access control[8];
- Policies, legal and ethical aspects[9].

Overall the user consultation allowed the working group to gather a valuable set of requirements for the development of the DMP common data model. Thus its greatest contribution was to help define the scope of maDMPs. There were obviously issues that still needed to be addressed. One of the issues that were detected, had to do with the granularity of the gathered information. Which proved to be too high for any meaningful analysis.

B. Second Consultation

As a result of the issues identified in the first user consultation, the decision to have a second user consultation was made. The objective was to narrow down the discussion, by focusing on specific fields and standards currently in use, as well as to define models for specific requirements. The target audience was stakeholders with expertise within the research data lifecycle.

This second user consultation took place at the RDA DMP Common Standards WG Workshop[10][11] hosted during the International Conference on Theory and Practice of Digital Libraries 2018 (TPDL 2018)[12].

Participants had to work in groups to solve two exercises. Each exercise was followed by a joint discussion.

1) First Excercise: In this exercise participants were divided into groups and asked to review a series of use cases displaying workflows that could potentially be automated under a maDMP ecosystem [18]. Thus, guiding researchers through the various stages of creating a DMP. The following 9 workflows were presented to the participants[13]):

- Start DMP;
- Specify Size and Type;
- Get Cost and Storage;
- Get Licence;
- Get Metadata Standard;
- Get Repository;
- Depoist Data;
- Get Help.

An example of one such use case is the *Get Cost and Storage* workflow, which deals with the selection, configuration, cost

[3]Github First Consultation Repository: https://github.com/RDA-DMP-Common/user-stories

[4]Full list of requirements: goo.gl/T9Coex

[5]Administrative, roles and responsibilities: goo.gl/hMqdk9

[6]Data: goo.gl/jtxBkZ;

[7]Infrastructure: goo.gl/cWbvs8

[8]Security, privacy and access control: goo.gl/wwSDP8

[9]Policies, legal, and ethical aspects: https://goo.gl/MiWEno

[10]Workshop website: http://rda-ws-tpdl2018.idsswh.sysresearch.org/

[11]Blog Post "Balancing theory and practice in research data management – innovation and opportunity in the Dublin Core community": https://goo.gl/SdJ6Zk

[12]TPDL 2018 Website: http://www.tpdl.eu/tpdl2018/

[13]A visual representation of these workflows as BPMN (Business Process Model and Notation) flows can be consulted here: https://goo.gl/72QFwX

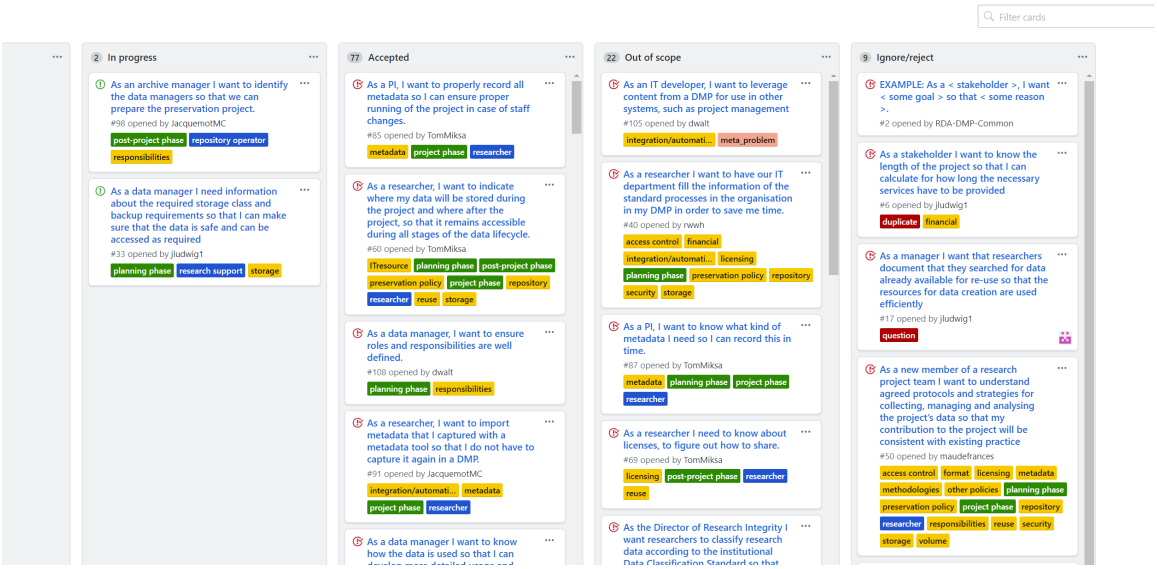


Fig. 2: GitHub project board used to organise user stories. User stories were sorted into three categories: accepted, out of scope or ignored/rejected. Additionally, labels that were assigned to user stories, as a means of classification can also be seen [16, p. 6].

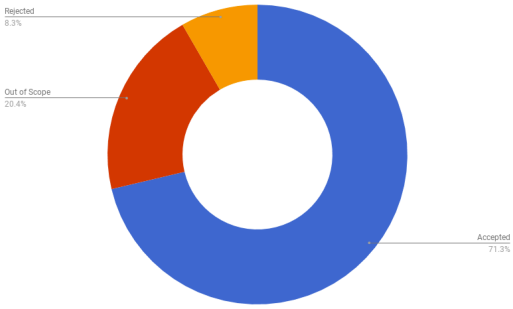


Fig. 3: Pie chart classification of collected user stories [16, p. 8].

estimation and provisioning of various kinds of storage used for managing research data during a project (active data). The workflow comprises of two process flows.

The first process flow, named *Storage Configuration and Cost Estimation* can be seen in Figure 4[14]. In this flow a researcher configures the required storage, that is offered by an information and communications technology (ICT) operator, and obtains a cost estimation for that storage.

The second process flow, named *Storage Provisioning* can be seen in Figure 5. This process takes place after the funding agency has approved the DMP/research project, and the necessary storage is both requested and provisioned.

These workflows show that data management planning requires inputs not only from researchers, but also from other stakeholders, such as infrastructure operators - only they can provide necessary details about quality of storage and costs. In order to successfully deploy a machine-actionable DMP at an institution, one needs to understand (and be able to model) similar interactions between stakeholders involved in data management.

[14]BPMN (This flow is depicted as a BPMN (Business Process Model and Notation: https://www.omg.org/spec/BPMN/2.0/) collaboration diagram

2) Second Exercise: **During the second exercise the participant groups were each assigned with documents containing sets of high level requirements. The documents pertained to each of the five categories first identified in the first user consultation (see section III-A for document links). Participants were asked to review the document they were assigned and to focus on three points:**

- **Further specify existing requirements;**
- **Provide examples of existing models, vocabularies or other sources, that could be reused in the DMP common data model;**
- **Provide both examples and justification of specific fields, that they would like to see introduced into the DMP common data model.**

Participants were then allowed to debate how to best reuse existing vocabularies and entities to specify the existing requirements. One of the most debated points, was how to express the various levels of data abstraction that are necessary to represent distinct stakeholders needs. To that effect, the concept of views was proposed.

The concept of views, focuses on the fact that distinct stakeholders have different needs. As such, the granularity of the information made available to them through the common DMP data model must reflect that. This concept can be illustrated by contrasting the needs of a repository manager with those of a researcher. A repository manager might require technical details about the research data that is to be stored in his repository, largely ignoring its content. On the other hand, a researcher might only be interested in analysing the generic content of the research data, with the technical details of such data being irrelevant to his concerns.

The second user consultation was attended by a total of 21 participants. Overall the second consultation allowed for narrowing down of the scope of the model and lead to definition of a concept of views.

IV. Tools to Automate Research Data Management

In this section we describe tools that demonstrate how research data management actions can be automated with a

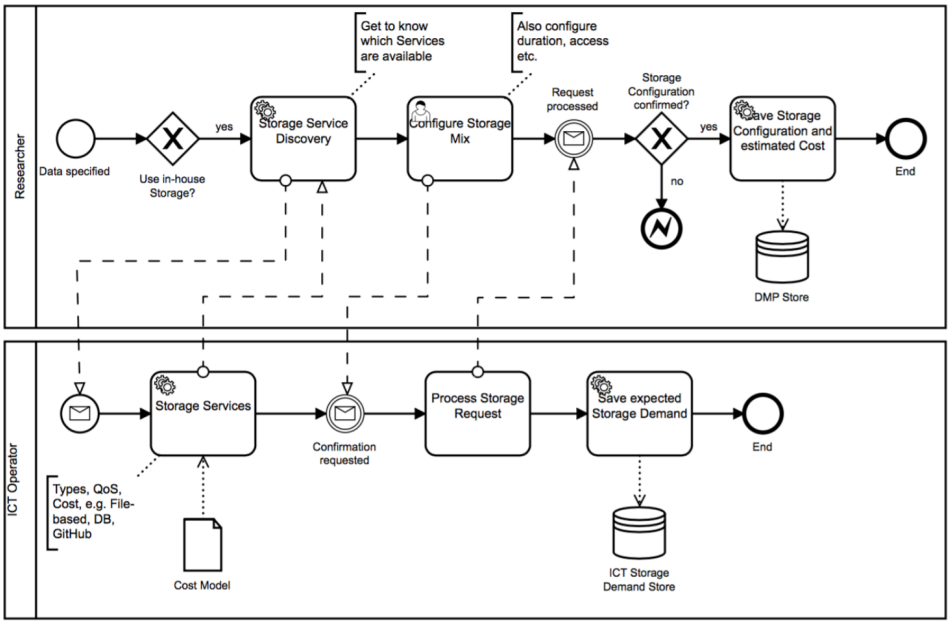


Fig. 4: Storage configuration and cost estimation.

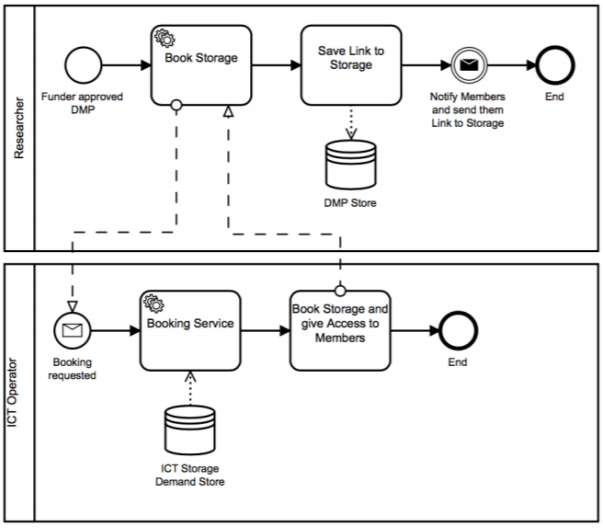


Fig. 5: Storage provisioning.

use of tools that read/write machine-actionable DMPs.

Students from the Technical University of Vienna (TU Wien)[15] enrolled for the course Digital Preservation[16] were asked to develop prototypes of automated DMP generator tools. The exercise[17] allowed students to choose the focus of their tool by selecting one of two main options:

1) **DMP creation on the onset of a research project. When funding has already been allocated and researchers need to create a DMP to describe information regarding the research data (i.e., quantifies of data that will be created, where will that data be stored, etc.);**
2) **DMP creation/update during the course, or at the close of a research project. When research data production is either underway or has already ceased, and researchers need to describe how information will be preserved.**

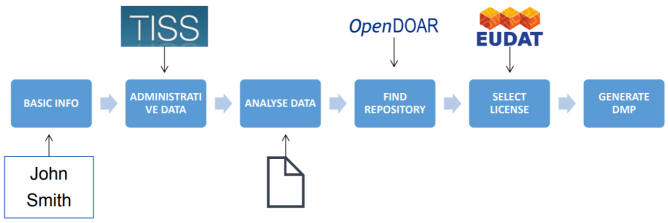


Fig. 6: DMP creation on the onset of a research project.

In the first option, the tool would provide storage estimates, and aid with the selection of an appropriate repository for research data storage – the overall objective of this option was to get estimations and recommendations. The process is described in Figure 6.

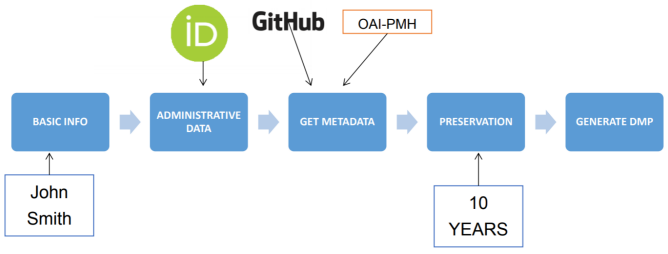


Fig. 7: DMP creation/updateDMP creation/update during the course, or at the close of a research project.

In the second option, tools had to retrieve information from third party services, to aid in the creation of a DMP or to update the DMP with real information by re-using information provided elsewhere. The process is described in Figure 7.

The common requirement for all tools was to require minimum input from users. For this reason, some of the

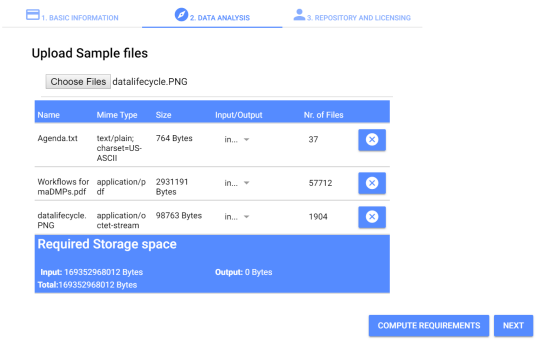
[15]TU Wien Website: https://www.tuwien.ac.at/
[16]Course Details: https://goo.gl/V6Zx3n
[17]Excercise: https://goo.gl/ZW3BFC

functionalities are limited. The goal was not to provide ready to use tools, but to demonstrate that a lot of information already exists in different systems and can easily be reused for a maDMP at different stages of research data life-cycle.

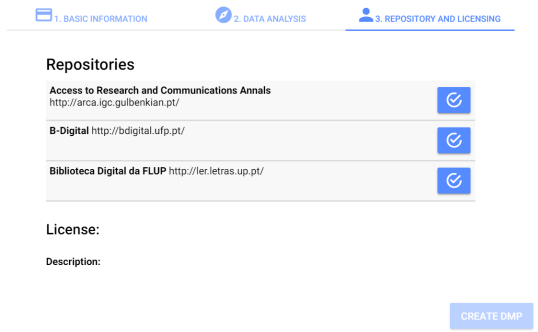
All three tools described in this section are featured in a blog post[18] by the DMPTool[19].

A. Planning phase: Tool 1

Tool 1 is an application that aids researchers in the creation of a DMP, during the planning phase of a research project[20]. This tool falls within option 1 (see Figure 6).



(a) Sample data.



(b) Repository suggestion.

Fig. 8: Tool 1 interaction examples.

The tool uses the Information Systems and Services of the TU Wien (TISS) API[21] to load user administrative data into a DMP. Users must then upload samples of data the research project will be producing. These samples are not expected to be real data, but mere representations of the data types that are to be produced.

As is seen in Figure 8a, users must indicate whether these data samples are to be considered input or output data. In addition to this they must also provide an estimate of number of files that are to expected for each of types of data they provided as samples. Tool automatically detects size and format of uploaded data. This information and minimal input required from researcher who annotates how many similar files will be used/produced helps in identifying storage needs.

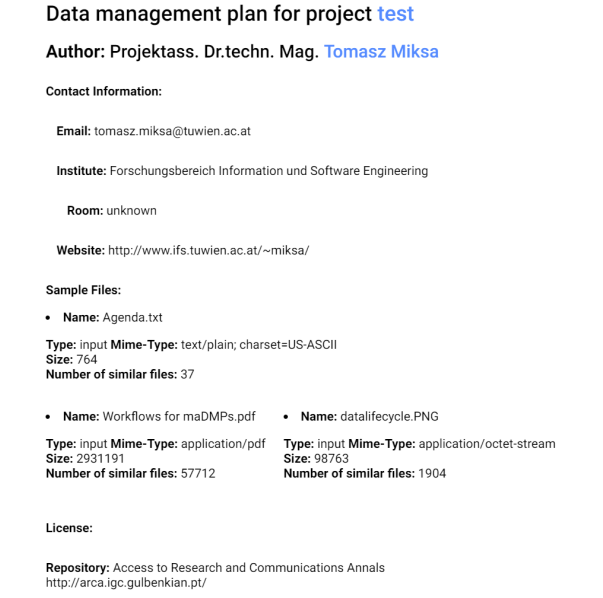
Based on the provided samples and inputs, Tool 1 is able to provide a suggestion of a repository that accepts the given data types and fulfilles further conidtions, e.g. must be located in Europe, is white listed by a funder, etc. (see Figure 8b).

[18]DMPTool Blog Post "Machine-actionable DMPs: What can we automate?" : https://goo.gl/ivhWkL

[19]DMPTool Website: https://dmptool.org/

[20]Tool 1 Github: https://github.com/IrinaAvram/DMPGenerator

[21]TISS: https://tiss.tuwien.ac.at/



(a) DMP administrative data.

Machine-actionable DMP


```
{
  "project": "test",
  "author": {
    "person": {
      "firstname": "Tomasz",
      "lastname": "Miksa",
      "precedingTitles": "Projektass. Dr.techn. Mag.",
      "postpositionedTitles": "",
      "mainEmail": "tomasz.miksa@tuwien.ac.at",
      "employee": {
        "employment": [
          {
            "organisationalUnit": {
              "value": "Forschungsbereich Information und Software Engineerin
            },
            "internalCode": null,
            "room": {
              "roomCode": "unknown"
            },
            "websites": {
              "website": [
                {
                  "value": "IFS website",
                  "url": "http://www.ifs.tuwien.ac.at/~miksa/"
                }
              ]
            }
          }
        ]
      }
    }
  },
  "repository": {
    "id": 1689,
    "name": "Access to Research and Communications Annals",
    "rUrl": "http://arca.igc.gulbenkian.pt/",
    "rOaiBaseUrl": "http://arca.igc.gulbenkian.pt/oaiextended/request"
  },
  "license": false,
  "files": [
    {
      "name": "Agenda.txt",
      "mimeType": "text/plain; charset=US-ASCII",
      "size": "764",
      "type": "input",
      "number": "37"
    },
    {
      "name": "Workflows for maDMPs.pdf",
      "mimeType": "application/pdf",
      "size": "2931191",
      "type": "input",
      "number": "57712"
    },
    {
      "name": "datalifecycle.PNG",
      "mimeType": "application/octet-stream",
      "size": "98763",
      "type": "input",
      "number": "1904"
    }
  ]
}
```


(b) Tool 1 maDMP.

Fig. 9: Tool 1 DMP examples.

Currently, recommendations are based by querying registries like openDOAR or re3data. Reducing the number of possibilities considerable eases selection of a correct repository. All of the information that is generated, is then persisted in a DMP. Figure 9a shows the human-readable version of a DMP, while Figure 9b shows the same information in a machine-actionable format.

With Tool 1 it was possible to demonstrate that:

- **Administrative information can be easily loaded into the DMP using existing interfaces;**
- **Non-technical users can be supported in estimating their storage needs;**
- **Repositories can be recommended automatically based on a set of criteria and by reusing existing APIs.**

B. Project- and post-project- phase: Tool 2

Tool 2 is an application that aids researchers in the creation of a DMP when the data has already been created and is either deposited in a repository accessible by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[22], or at Github[23]. This tool falls within option 2 (see Figure 7).

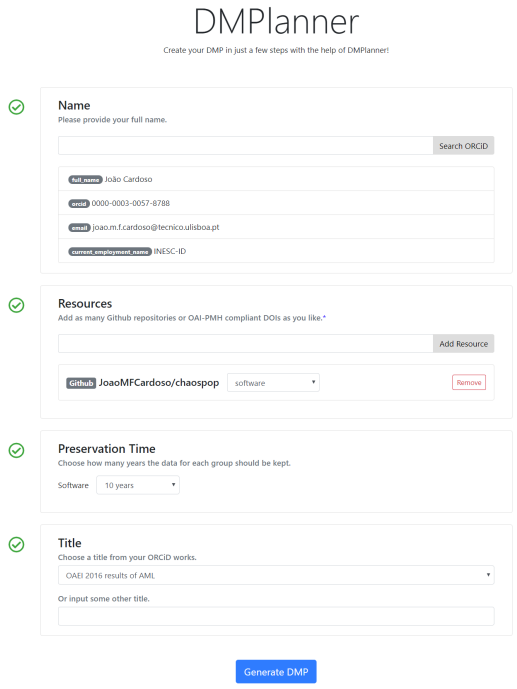


Fig. 10: Tool 2 interaction.

In Tool 2, users are asked to provide information to fill 4 fields: name, resources, preservation time and title. This process is illustrated in Figure 10.

In the *Name* field, users must provide a name associated with an ORCID[24] profile. So that the application can import administrative data on the user. In the *Resources* field, users should provide at least one resource Digital Object Identifier (DOI)[25]. This DOI can resolve to either a Github or OAI-PMH compliant repository. They must then classify all added resources according to their type (e.g., documentation, software,

[22]OAI-PMH Website: https://www.openarchives.org/pmh/

[23]Tool 2 Github: https://github.com/alexschwarzresearch/DMPlanner

[24]ORCID website: https://orcid.org/

[25]ISO 26324:2012: https://www.iso.org/obp/ui/#iso:std:iso:26324:ed-1:v1:en

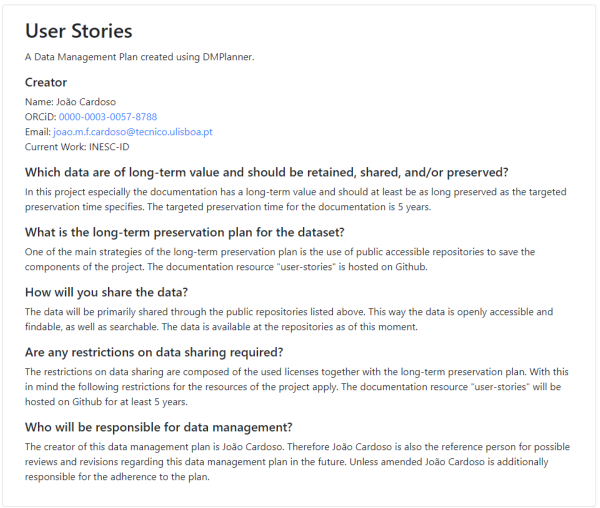


Fig. 11: Tool 2 generated DMPs.

presentation, etc.). In the *Preservation Time* field, users are asked to specify the amount of time in years they wish to preserve each of the resource types. Finally in the *Title* field, users are required to input a title or choose one from one of the works associated with the provided ORCID profile.

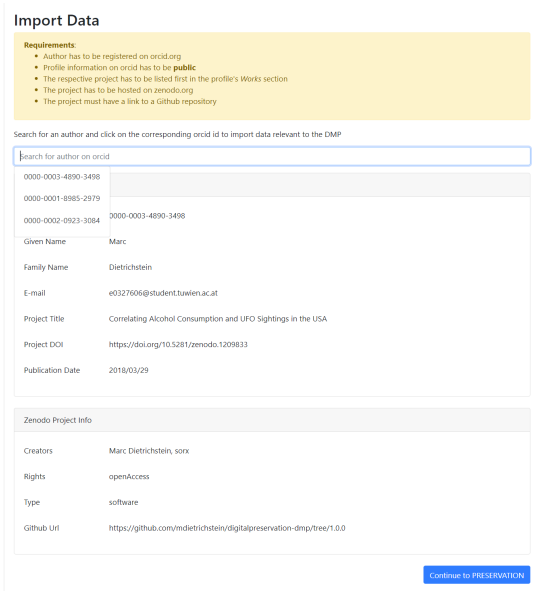
Once all 4 fields are correctly filled, users are allowed to generate a DMP. Tool 3 will then create two artefacts: a human readable DMP that follows a predefined template, and a maDMP, as seen in Figure 11.

With Tool 2 it was possible to prove that:

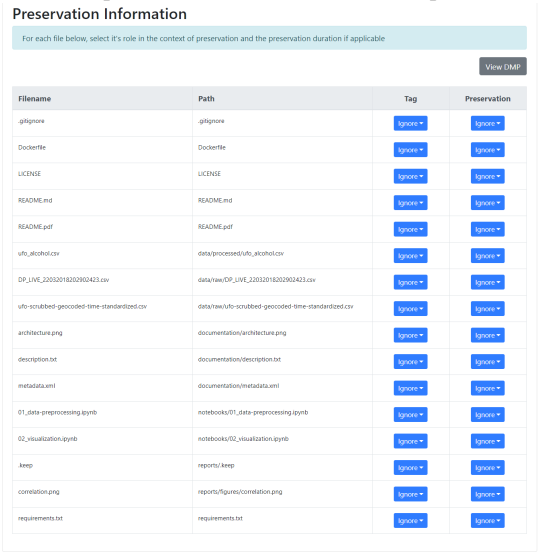
- **Persistent identifiers, such as DOI, can facilitate generation of a DMP;**
- **Information needed for a DMP can be imported from external systems, thus eliminating the need re-input existing information;**
- **Automatically generated human-readable DMPs can be of high quality.**

C. Project- and post-project- phase: Tool 3

Tool 3 is also an application to aid researchers in creating or updating a DMP based on existing information[26]. This tool also falls within option 2 (see Figure 7).



(a) Imported data based on an ORCID profile.

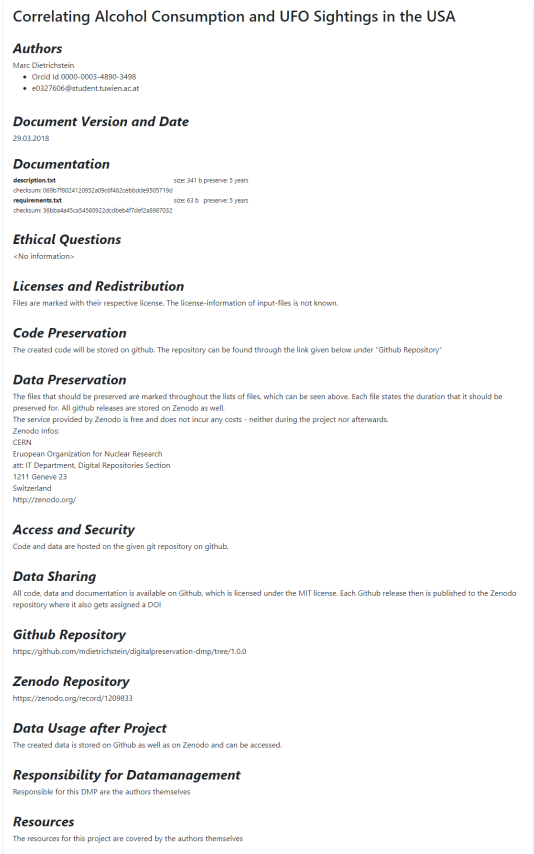


(b) List of available files within the repository associated with the ORCID profile.

Fig. 12: Tool 3 interaction examples.

With Tool 3 users are asked to provide an ORCID profile. The administrative information contained in the profile is imported, as well as any recent works associated with the profile, as can be seen in Figure 12a. The following step assumes that users have previously assigned a DOI to a Github repository, and then proceeded to include that repository in their list of ORCID projects. If that condition is met, users are presented with a list of all the available files within the ORCID associated repositories. Users are then asked to provide classification information regarding those files by attributing

[26]Tool 3 Github: https://github.com/mdietrichstein/digitalpreservation-dmp-generator



(a) Human readable DMP.

(b) Tool 3 maDMP.

Fig. 13: Tool 3 DMP examples.

tags to each of the files (e.g., ignore, input data, documentation, publication, data, etc.). Users must also classify the files in accordance to their preservation policy, i.e., how long should they be preserved, if at all. An example of a list of files and their respective classifications, can be seen in Figure 12b. The user can then order for a DMP to be generated. Tool 3 will then generate two versions of the DMP. One aimed at human consumption (see Figure 13a) and a machine-actionable version (see Figure 13b).

With Tool 3 it was possible to prove that:

- Administrative data can be imported from any Current Research Information System (CRIS);
- Services for storing data can provide relevant information on how data is managed;
- An automatically generated DMP can have higher granularity of information, e.g.:
  - All types of data are listed (researchers often focus solely on input or output data disregarding other data types);
  - Each file can have a hash that allows for both its identification and compliance validation;
- Free form text cannot be avoided in an maDMP.

V. Conclusions

In this paper we presented ongoing efforts to develop machine-actionable Data Management Plans that allow systems to act on behalf of researchers and other stakeholders involved in research data management. Thus, reducing workload put on researchers by automating planning and actual management of data.

We described two consultations performed by the RDA DMP Common Standards working group.

The first consultation, aimed at defining the scope of the common data model for machine-actionable Data Management Plans. It helped to define the concept of "machine-actionability" in the context of Scientific Data Management, and allowed the community to contribute with examples of user stories that describe specific challenges and requirements that should be addressed by a machine-actionable DMP.

The second consultation reached out to experts with domain knowledge. The objective was to break down collected requirements into specific fields and to identify models and concepts that can be reused in creating a common data model for a machine-actionable DMP. Thus, the second consultation allowed to narrow down the focus on a core set of requirements, and paved the way to establish a clear structure of the model.

In this paper we also presented three proof of concept tools that demonstrate how researchers can benefit from a machine-actionable DMP when the common data model and necessary services are in place. The presented tools show how existing information stored in various systems can be reused to either provide estimates and recommendations in a planning phase of a project, e.g. to estimate costs, or to provide real values during the project, e.g. recommend repositories, or provide information on licenses.

The next steps for the DMP Common Standard working group is to turn results of consultations, feedback and experience collected when developing proof of concept tools into the common data model for a machine-actionable DMP. During the 12th RDA plenary meeting[27] the group will review a first draft of the common data model and will launch a third open consultation that will specifically focus on:

- establishing the set of classes and properties that make up the core data model, by reviewing its first draft;
- presenting the concept of DMP components, that should be viewed as extensions to the core data model;
- defining a set of guidelines on how to both customise and implement the DMP common data model.

[27]RDA 12th plenary meeting: https://goo.gl/Spvrq4

Acknowledgments

This work would not be possible without support of enthusiastic RDA DMP Common Standards group members and its chairs: Paulk Walk and Peter Neish. The authors would like to thank to: Irina Avram, Marc Dietrichsten, Victor Dulca, Markus Neumeyer, Simon Oblasser, and Alex Schwarz who implemented proof of concept tools.

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, and by project PRECISE, Accelerating progress toward the new era of precision medicine, 2016-2019 (LISBOA-01-0145-FEDER-016394). This research was also carried out in the context of the Austrian COMET K1 program and publicly funded by the Austrian Research Promotion Agency (FFG) and the Vienna Business Agency (WAW).

References

[1] J. Gray, D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt, and G. Heber, "Scientific data management in the coming decade," *Acm Sigmod Record*, vol. 34, no. 4, pp. 34–41, 2005.

[2] T. Hey and A. Trefethen, "The data deluge: An e-science perspective," *Grid computing: Making the global infrastructure a reality*, pp. 809–824, 2003.

[3] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen *et al.*, "Promoting an open research culture," *Science*, vol. 348, no. 6242, pp. 1422–1425, 2015.

[4] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, 2016.

[5] Q. Schiermeier, "Data management made simple," *Nature*, vol. 555, no. 7696, pp. 403–405, 2018.

[6] A. Surkis and K. Read, "Research data management," *Journal of the Medical Library Association: JMLA*, vol. 103, no. 3, p. 154, 2015.

[7] W. K. Michener, "Ten simple rules for creating a good data management plan," *PLoS computational biology*, vol. 11, no. 10, p. e1004525, 2015.

[8] S. Simms, S. Jones, D. Mietchen, and T. Miksa, "Machine-actionable data management plans (madmps)," *Research Ideas and Outcomes*, vol. 3, p. e13086, 2017. [Online]. Available: https://doi.org/10.3897/rio.3.e13086

[9] J. Gray, *TJim Gray on eScience: A transformed scientific method*. Microsoft research Redmond, WA, 2009, vol. 1, pp. 5–12.

[10] A. Whyte and J. Tedds, "Making the case for research data management," *DCC Briefing Papers*, 2011. [Online]. Available: http://www.dcc.ac.uk/resources/briefing-papers

[11] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, "Addressing big data issues in scientific data infrastructure," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*. IEEE, 2013, pp. 48–55.

[12] J. Qin, "Infrastructure, standards, and policies for research data management," 2013.

[13] T. Miksa, S. Simms, D. Mietchen, and S. Jones, "Ten simple rules for machine-actionable data management plans (preprint)," Feb. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1172673

[14] T. Miksa, P. Neish, and P. Walk, "Wg dmp common standards case statement," 2017. [Online]. Available: https://www.rd-alliance.org/group/dmp-common-standards-wg/case-statement/rda-wg-dmp-common-standards-case-statement

[15] T. Miksa, R. Vieira, J. Barateiro, and A. Rauber, "Vplan–ontology for collection of process verification data," 2014.

[16] T. Miksa, P. Neish, P. Walk, and A. Rauber, "Defining requirements for machine-actionable data management plans," 2018. [Online]. Available: http://ifs.tuwien.ac.at/~miksa/papers/2018-iPres-maDMPs.pdf

[17] **M. Cohn,** ***User stories applied: For agile software development*****. Addison-Wesley Professional, 2004.**

[18] **A. Bakos, T. Miksa, and A. Rauber, "Research data preservation using process engines and machine-actionable data management plans," in** ***Digital Libraries for Open Knowledge, 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Portugal, September 10-13, 2018, Proceedings.*****, ser. Lecture Notes in Computer Science, E. Méndez, F. Crestani, C. Ribeiro, G. David, and J. C. Lopes, Eds., vol. 11057. Springer, 2018, pp. 69–80. [Online]. Available: https://doi.org/10.1007/978-3-030-00066-0_6**