Defining requirements for machine-actionable Data Management Plans

Tomasz Miksa
SBA Research & TU Wien
Favoritenstrasse 16
1040, Wien
tmiksa@sba-research.org

Peter Neish
Digital Scholarship, Research and Collections
The University of Melbourne,
3010, Australia
peter.neish@unimelb.edu.au

Paul Walk
Antleaf Ltd.
Ground Floor Unit B Lostock
Office Park Lynstock Way,
Lostock, Bolton
England, BL6 4SG
paul@paulwalk.net

Andreas Rauber
TU Wien
Favoritenstrasse 16
1040, Wien
rauber@ifs.tuwien.ac.at

ABSTRACT

Data Management Plans (DMPs) are free-form text documents describing data used and produced in research projects. The workload and bureaucracy often associated with traditional DMPs can be reduced when they become machine-actionable. However, there is no common definition of what machine-actionable DMPs really are. This hinders the communication between stakeholders and leads to scepticism, or conversely to exaggerated expectations.

This paper aims to clarify what machine-actionable DMPs are and provides examples of how involved stakeholders can benefit from them. It describes an open stakeholder consultation performed by the RDA DMP Common Standards working group. The main objective was to define the scope of information covered by machine-actionable DMPs and formulate an initial set of requirements for a common data model for machine actionable DMPs. To do this we used methodology known from system and software requirements engineering to collect information on how needs for information of particular stakeholders evolve over phases of the research data lifecycle.

Keywords

DMPs, maDMPs, machine-actionable, data management, rda, common data model

1. INTRODUCTION

Data Management Plans are documents accompanying 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. The current manifestation of a DMP - a static document often created before a project begins - only contributes to the perception that DMPs are an annoying administrative exercise and do not support data management activities. Questions can remain unanswered, or the answers can be overly generic due to the use of free-form text.

What DMPs really are - or at least should be - is an integral part of research practice, since today most research across all disciplines involves data, code, and other digital components. We continue to need a human-readable narrative, but there is now widespread recognition that the DMP could have more thematic, machine-actionable richness with added value for all stakeholders. This includes researchers, funders, repository managers, administrators, data librarians, and so on; in short, everyone who is part of the larger ecosystem in which data is produced, transformed, exchanged, reused, and preserved.

To achieve this goal, all stakeholders must coordinate efforts to realize a new generation of machine-actionable DMPs (maDMPs) that contain an inventory of key information about a project and its outputs. The basic framework requires common data models for exchanging information, as well as a shared ecosystem of services that send notifications and act on behalf of humans to collect the necessary information in a semi-automatic way. Thus reducing the actual workload on everyone involved in the research data lifecycle.

However, there is no common definition of what machine-actionable DMPs really are. There are many misconceptions and misunderstandings among stakeholders. This, in turn, leads to scepticism towards the maDMPs, because the stakeholders are concerned that maDMPs not only will not resolve the problems of traditional DMPs, but also will increase the workload and bureaucracy.

In this paper we aim to clarify what machine-actionable DMPs really are and provide examples of how involved stakeholders can benefit from them. We also describe an open stakeholder consultation performed by the RDA DMP Com-

iPres '18 Boston, Massachusetts USA

mon Standards working group which had the main objective of defining the scope of information covered by machine-actionable DMPs. Thus, we clarify their definition and facilitate the discussion among interested parties. For that purpose we used a methodology from the domain of software and system requirements engineering. We used so-called user stories that express in a natural language stakeholders' requirements towards maDMPs. To facilitate their collection and processing we used GitHub and organised workshops. We managed to reach out to stakeholders from Africa, Australia and Europe. As a result, we identified: (1) which stakeholder groups are involved in processing maDMPs; (2) which stakeholders need information from maDMPs; (3) which stakeholders can provide this information. Based on these findings, we defined an initial set of requirements for a common data model for machine-actionable DMPs.

The paper is organised as follows. Section 2 discusses related work on DMPs. Section 3 clarifies what machine-actionability in view of DMPs means and provides examples of how maDMPs can change traditional research data management and preservation. Section 4 describes methodology of the consultation. Section 5 discusses results of the consultation. Conclusion and future work are provided in Section 6.

2. RELATED WORK

Data Management Plans (DMPs) are documents accompanying research proposals and projects. They 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, and the National Research Foundation (NRF) in South Africa.

DMPs are created usually using checklists [1] or templates provided by funders, e.g. Horizon 2020 template that requires data to be FAIR - Findable, Accessible, Interoperable and Reusable [12]. There are also a number of tools like DMP Online[1], DMP Tool[2], or RDM Organizer [2] which provide templates and tailored guidance based on specific funder requirements. Most of the DMP templates have similar sections and for this reason general rules on how to create them apply [7].

Automation and machine-actionability were identified as key factors enabling deployment of the European Open Science Cloud (EOSC) [4]. The EOSC aims to create a trusted environment for hosting and processing research data to support EU science in its global leading role. This leads to creation of a European data economy and is part of the Digital Single Market strategy of the European commission.

Machine-actionability and DMPs are also in the focus of the Research Data Alliance[3] (RDA) which is an international organization focused on the development of infrastructure and community activities aimed to reduce barriers to data sharing and exchange, and promote the acceleration of data driven innovation worldwide. RDA established several working groups that deal with management, sharing, and preservation of data.

One of them is the DMP Common Standards[4] working group that works to realise a vision where DMPs are developed and maintained in such a way that they are fully integrated into the systems and workflows of the wider research data management environment. To achieve this vision the group is developing a common data model with a core set of elements. Its modular design should allow customisations and extensions using existing standards and vocabularies to follow best practices developed in various research communities [11]. The data model will use semantic technologies which were successfully applied in the domains of data management and preservation [9]. The common data model is NOT intended to be a prescriptive template or a questionnaire, but to provide a re-usable way of representing machine-actionable information on themes covered by DMPs.

The need for establishing DMP Common Standards working group was articulated during the 9th plenary meeting in Barcelona during the Active DMPs IG session. The discussion was framed by a white paper [10] on machine-actionable data management plans (DMPs). The white paper is based on outputs from the IDCC workshop held in Edinburgh in 2017 that gathered almost 50 participants from Africa, America, Australia, and Europe. It describes eight community use cases which articulate consensus about the need for a common standard for machine-actionable DMPs (where machine actionable is defined as "information that is structured in a consistent way so that machines, or computers, can be programmed against the structure"[5])

There are a number of technical platforms with potential applications for machine-actionable DMPs. The summary of ongoing initiatives and tools can be found in the *activedmps.org* portal.

3. MACHINE-ACTIONABLE DMPS

In this section we establish context for the requirements engineering exercise described in Section 3. First, we explain the difference between machine-actionable DMPs and traditional DMPs. Second, we demonstrate which stakeholders are involved and how they can collaborate using maDMPs.

3.1 Data Model

Traditional DMPs are supposed to be living documents that describe how data will be managed during the life of a project. DMPs should state what data will be created and how, as well as outline plans for sharing and preserving data.

In most cases DMPs are created using templates or online questionnaires such as DMP Online or DMP Tool. They can be later exported to a range of formats such as PDF or DOCX, but also to XML or JSON. The first are to be used by humans, the latter are predominantly intended to be used by machines. However, using a machine-actionable file format does not make DMPs itself machine-actionable.

The real challenge is to model the information provided in a DMP in a machine-actionable way. To illustrate the challenge we have modelled an excerpt of a DMP that describes basic administrative data about a principle investigator who created a DMP. This is depicted in Figure 1.

The upper part of the figure shows a typical DMP that was exported to XML. There is a question and an answer to it. Both the question and the answer are provided using free-form text. The XML tags let us (and the machines)

[1] https://dmponline.dcc.ac.uk
[2] https://dmptool.org
[3] https://www.rd-alliance.org
[4] https://www.rd-alliance.org/groups/dmp-common-standards-wg
[5] http://www.ddialliance.org/taxonomy/term/198

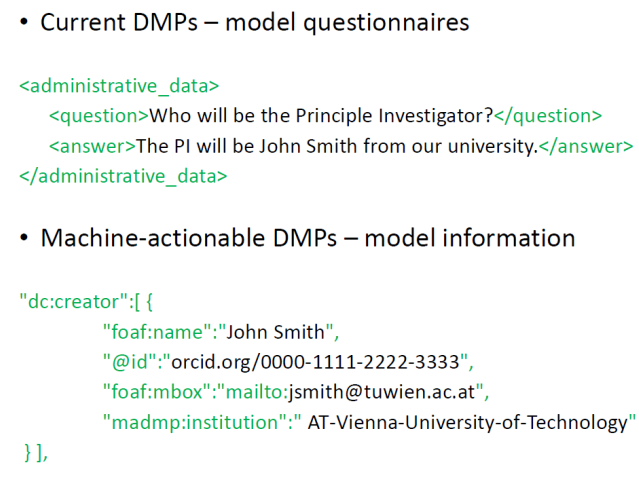

Figure 1: Comparison of models for traditional DMPs (upper part) and machine-actionable DMPs (lower part).

distinguish between questions and answers. Apart from this simple distinction, machines are not able to learn anything about the content of the DMP without using complex text analysis algorithms. Such a representation is not machine-actionable.

The lower part of the figure shows how the same information can be modelled in a machine-actionable way. The free-form text was replaced with specific fields, such as: name, id, mail, institution. Thus, additional semantics have been added. This allows queries to be built which can be run by machines to source information from DMPs, for example, to identify the principle investigator of a DMP.

This was possible, because we modelled the information about the principle investigator using controlled vocabularies and existing standards, for example, Dublin Core. We did not model the structure of the questionnaire which was used to facilitate the DMP creation. Information contained in a DMP must be useful to other stakeholders such as repository operators and data librarians. It is not important how the DMP was created, but what information it contains.

Furthermore, when information is not coupled with a specific question, then it can be used for other purposes, for example, John Smith used as an example in Figure 1 not only can be a principle investigator, but also a person responsible for implementing the DMP. In such case, it is enough to add another field describing his role. In a traditional DMP we would have a new question and very likely identical answer that was copy pasted.

3.2 Stakeholders and the ecosystem

A data model that models real information, not just a questionnaire structure, is only the first step towards machine-actionability. We also need an ecosystem of services that read and write information from and to DMPs on behalf of various stakeholders, such as, repository operators, funders, legal experts, etc.

In most cases researchers are those who are solely responsible for the contents of a traditional DMP. They are encouraged to contact IT experts or repository operators when writing their DMPs. Researchers often do not know whom to ask, or it is too late to prepare a good quality DMP. Many misunderstandings and mistakes can be avoided if the right stakeholders are contacted at the right time. For instance, if a researcher is planning an experiment producing big data, then a storage operator should confirm that the produced data can be maintained during the project. Furthermore, he should provide information on costs so that the researcher or the reviewer can validate if the costs can be covered by the project. Otherwise, it may happen that a well-designed experiment will fail, because the DMP and the proposal did not include real data management costs. Such problems can be avoided if all stakeholders are involved at the right time in the process of DMP creation.

Machine-actionable DMPs can help solve such challenges allowing stakeholders to exchange information using them. Dedicated services can act on behalf of stakeholders, for example, by providing information on costs of data management for a requested amount and type storage. Such services can also automate other tasks and can send notifications to stakeholders depending on a state of a machine-actionable DMP. Below we discuss two examples of how the ecosystem of services acting on behalf of stakeholders can be used together with machine-actionable DMPs

Figure 2 describes a typical process for creating a DMP in the initial phase of a project. Depending on a funder policy this happens either before the project starts, for example NSF in the US, or within the first few months after the project has started, for example Horizon 2020 in Europe. In a traditional scenario researchers would have to provide all information, such as, file types, costs of storage, licenses to be used, etc. on their own. In the presented scenario we show how the workload related to this traditional scenario can be reduced by involving other stakeholders and using services acting on their behalf.

First a researcher starts a new DMP. Administrative information such as affiliation is automatically imported from institutional employees' database or ORCID. Thus, the researchers do not have to provide yet again the same information which already exists in a different system. In the next step researcher specifies size and type of data that will be created in the project. Based on this basic information, the next three steps can be executed automatically using following services:

- storage booking — a service that acts on behalf of an infrastructure operator and reserves storage space for the duration of a project if a repository suitable for the expected types and amounts of data and meeting relevant policy requirements can be found. Furthermore, such a service can help repository managers plan investments into infrastructure when knowing in advance how much new data is expected within the planning period.

- cost estimation — a service that acts on behalf of repository operators and implements a cost model of a repository to provide automatic estimates of costs of storage and preservation based on input parameters such as amount of data, type of data, project duration, etc. There has been research on cost models and ways of comparing them [6], but there is still no such service in place.

- license selection — a service that acts on behalf of legal

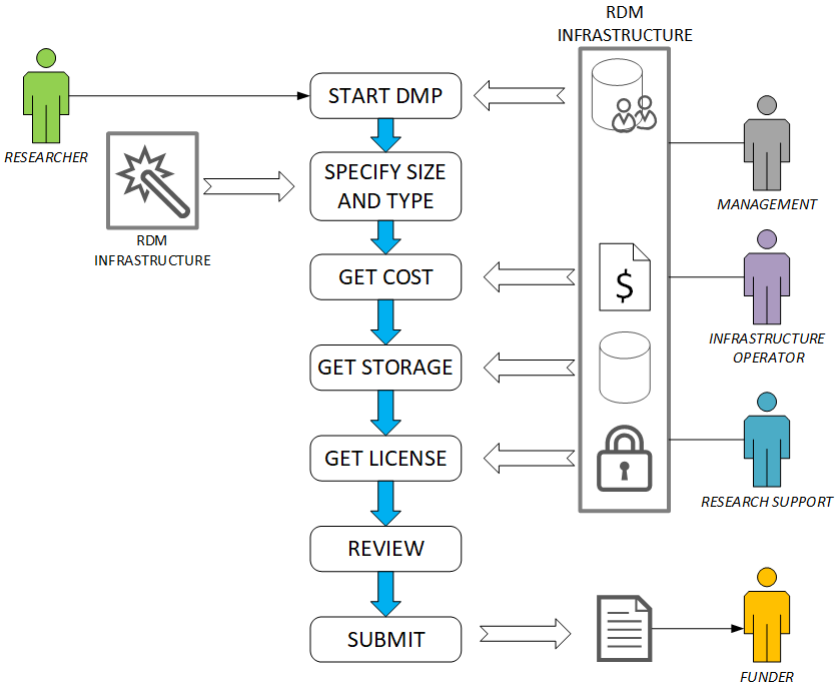

Figure 2: Typical process for creating a DMP in the initial phase of a project. It shows various stakeholders that are involved/affected by decisions taken at different steps.

experts and proposes a license for data sharing, taking into account policies that apply to the project and type of data. For example, if the institutional policy recommends open access publishing and the data do not contain sensitive information, then CC0 could be the default setting for data, and CC BY for text and media. There is already a wizard from EUDAT [3] that offers similar functionality.

In the next step, the researcher can review the contents of generated parts and adapt them if necessary. Finally, the DMP is submitted to the funder who requires the DMP.

This example shows that effort required to create maDMPs is lower compared to the traditional DMPs. Due to automation, the proposed approach does not increase bureaucracy and does not overload other stakeholders with new tasks.

Another example of stakeholder cooperation is depicted in Figure 3. It demonstrates how stakeholders communicate with each other by exchanging information through DMPs at a later stage of a project, that is, when data exists and is being preserved.

Using information from the maDMP a repository operator can select a proper repository, set an embargo period, and assign a correct license to data submitted by researchers. In return, system acting on behalf of a repository operator provides a list of DOIs assigned to the data and provides information on costs of preservation. This in turn can be accessed by a funder to check how the DMP was implemented. Researchers can browse DMP catalogues using a variety of filters that allows them to discover projects using similar methodologies or infrastructure or producing similar outputs.

There are many more use cases in which maDMPs and supporting services can orchestrate exchange of information between stakeholders involved in data management and preservation. We picked these two to illustrate what machine-actionability means in view of data management plans and to stimulate the discussion on what the community expects from maDMPs.

4. REQUIREMENTS ENGINEERING

In this section we describe motivation, goals and methodology used to identify requirements of stakeholders towards machine-actionable DMPs.

4.1 Goals

We observed during discussions with interested stakeholders that there is still no clear and common understanding of what a machine-actionable DMP is or which specific information it should contain. This is because the stakeholders have different backgrounds and the term is used in various, but related contexts. For example, the European Commission sees the machine-actionability as one of the enablers for the European Open Science Cloud.

Checklists and templates developed for traditional DMPs influence the way people think about machine-actionable DMPs. The traditional DMPs are limited to information that can be provided manually in a reasonable amount of time usually by researchers. Machine-actionable DMPs do not have this limitation, because information can be sourced automatically and can be more detailed. As a result, information that was not covered previously can be modelled now. For this reason we aim to identify which information should be part of maDMPs.

The maDMPs can only become living documents when they are able to address needs of stakeholders involved in the data research lifecycle. The needs of stakeholders change

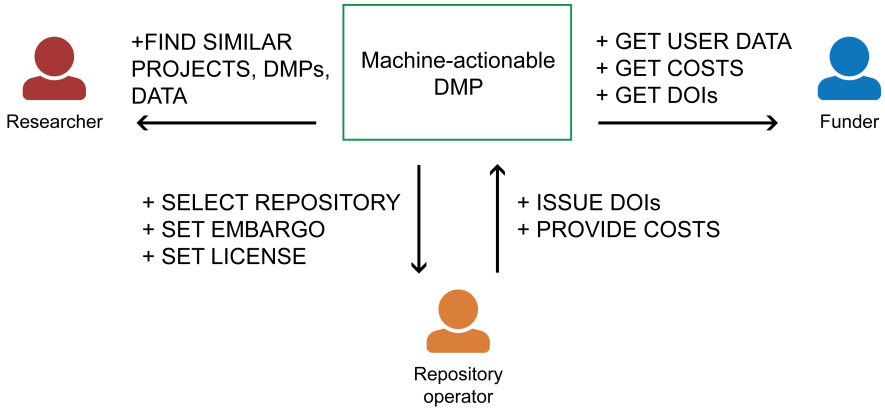

Figure 3: Example of a seamless information exchange between stakeholders through machine-actionable DMPs.

over time. An initial DMP created before project starts can provide only estimations, while a DMP created when project finishes should describe existing data and provide details on preservation. Hence, we aim to identify who needs which information and when, and who can provide this information and when.

The consultation and requirements analysis attempts to find answers to following questions:

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

The results are used by the RDA DMP Common Standards working group to develop a common data model for machine-actionable DMPs (cf. Section 2).

4.2 Methodology

In the course of work of the RDA DMP Common Standards working group we performed a user consultation to answer goals set in Section 4.1.

4.2.1 *Collecting user stories*

We used the GitHub to enable the community to contribute suggestions and resources in the open. We asked participants to submit their ideas using the issue mechanism provided by the GitHub. Thus, participants had a possibility not only to review and comment what others had contributed, but also to define new user stories by raising new issues.

User stories are used in software development to describe features of software or systems. They are written using natural language and express expectations of end users. The user stories can be later broken down into specific functional and non-functional requirements by system/software architects. This approach was applicable in case of our consultation, because our aim was to reach out to a wide range of stakeholders and we had to provide an easy way for them to provide their feedback. Furthermore, to structure and ease formulation of user stories we provided a template:

As a <stakeholder>, I want <goal> so that <reason>.

We also provided examples such as:

- *As a researcher, I want to inform repository operator on the amount of data in the planning phase, so that they provide information on costs.*
- *As a repository operator, I want to know the embargo periods for ingested data, so that I can restrict access to specific contents and comply with policies.*

The call for contributions was open between 09 October 2017 and 30 November 2017. We shared the invitation for consultation through mailing lists and twitter. Furthermore, we organised workshops to reach out to stakeholders not directly involved in the RDA activities [8]. User stories collected in workshops were put online, usually by a single user. For this reason it may appear that only few distinct users contributed to the study, but this is not the case. All collected user stories can be found online[6].

4.2.2 *Organising and classifying user stories*

We used the GitHub project board to review and classify user stories using labels (see Figure 4 and the full version online[7]). This helped use to get an overview of collected user stories. We used six columns to organise the issues: *unclassified*, *read for review*, *in progress*, *accepted*, *out of scope*, *ignore/reject*. We moved the issues (user stories) between the columns as we progressed with their classification.

We used labels to classify the user stories. The labels helped us divide user stories by: stakeholders, project phases, and subject of information conveyed in a DMP.

Labels referring to stakeholders used a blue colour and included following stakeholders: *researchers*, *funders*, *repository operators*, *service providers* (services other than the repository holding the data), *research support*, *institutions*. We used these labels to indicate to which of them a user story refers to.

The green labels referred to research data lifecycle phases. We analysed existing research data lifecycle models such as the DCC curation lifecycle model [5] or the model created by the University of Central Florida[8], but all of them defined

[6] `https://github.com/RDA-DMP-Common/user-stories/`

[7] `https://github.com/RDA-DMP-Common/user-stories/projects/2`

[8] `http://guides.ucf.edu/ScholarlyCommunication/ResearchLifecycle`

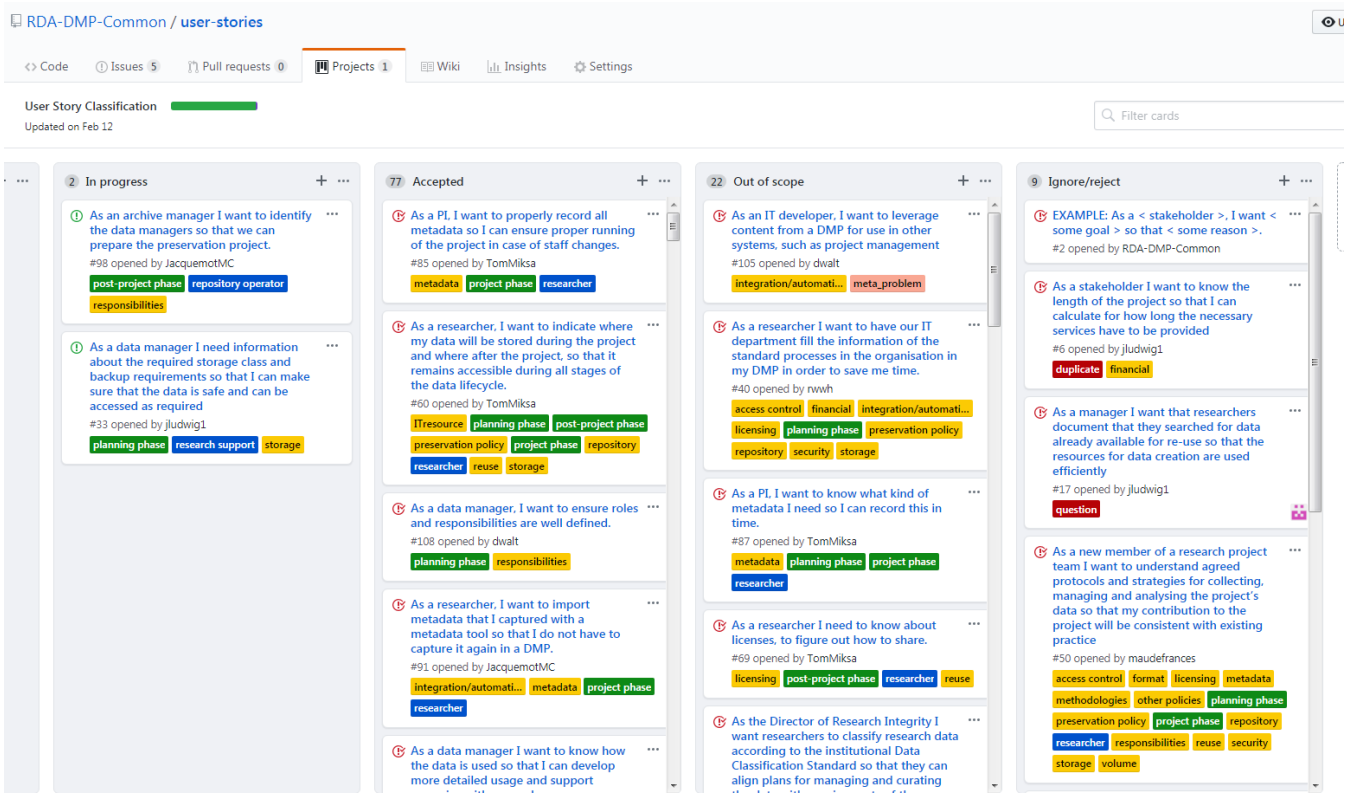

Figure 4: GitHub project board used to organise user stories. It depicts which user stories were accepted and which are out of scope. Furthermore, one can see labels that were assigned to each of the user stories to classify them.

phases that are not self-explanatory, that is, they cannot be understood correctly by people who are not familiar with the rest of the model. For this reason, we simplified the research data lifecycle to three phases:

- planning phase — before the actual research starts, e.g. when writing a grant proposal or a provisional DMP,
- project phase — when the research is performed, results are being published, etc.,
- post-project phase — when the project has finished and data has to be shared and preserved.

The last group of labels were yellow labels that we used to indicate the subject of the information being conveyed in the DMP. They are inspired by the DMP Roadmap themes[9] and include labels such as: *volume*, *metadata*, *licensing*, *security*, etc. A full list of labels can be found online[10].

For example, user story #101[11] states: "As an archive manager, I want to know in advance the conservation period of data so that I can better organize the service and adapt the preservation actions." We identified that it expresses requirements of *repository operators* in both *planning* and *project phases*, and deals with *preservation policy*. Thus, using filtering options provided by the GitHub, we were able to identify all user stories referring to a particular stakeholder, project phases, or a DMP theme. This helped navigation among user stories and identifying which topics or stakeholder groups are better represented in the consultation than others.

4.2.3 *Visualising user stories*

We visualised the issues and their labels to discover dependencies between specific themes, phases and stakeholders – see Figure 5.

The visualisation shown in Figure 5 allowed us to quickly see which were the most popular stakeholders, subjects and phases, and how strongly associated they were. The visualisation is interactive where any of the stakeholders, subjects or phases can be selected and the relevant associations displayed. The thicker the line the stronger the relationship.

In Figure 5 the *preservation policy* subject has been selected and all the related labels are shown. We can see that information on preservation policy is needed in all three phases of research data lifecycle, but the thickest line indicates that more user stories identified the *planning* phase as important. Furthermore, according to the user stories only *institutions* and *repository operators* need to know about

[9] https://github.com/DMPRoadmap/roadmap/wiki/Themes

[10] https://github.com/RDA-DMP-Common/user-stories/wiki

[11] https://github.com/RDA-DMP-Common/user-stories/issues/101

[13] https://bl.ocks.org/peterneish/f6dad14e46327011f0ccf15d49dd27fb

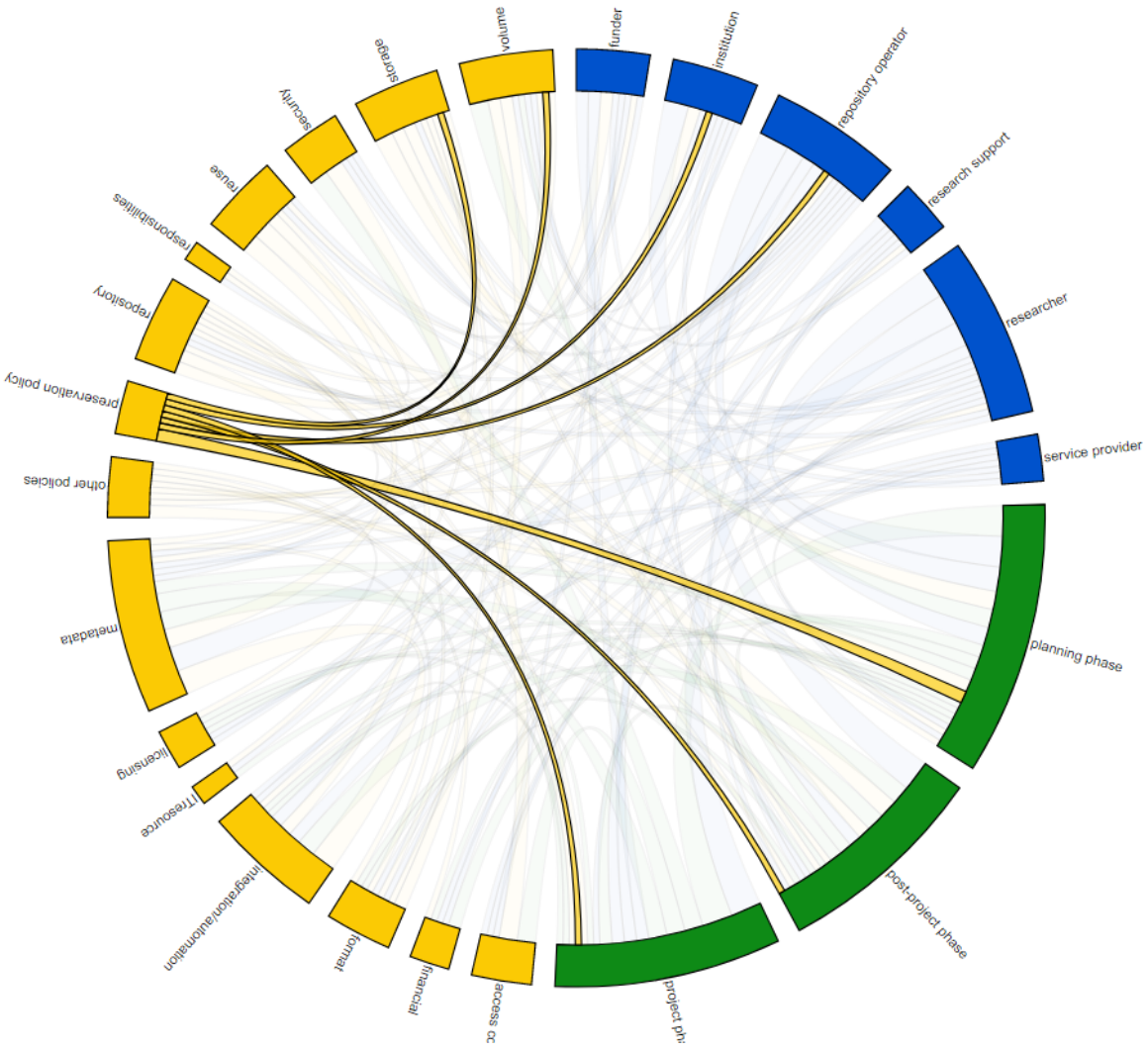

Figure 5: Visualisation of collected user stories showing connection between stakeholders, phases, and DMP themes. The example presented in the figure highlights one selected option. A full interactive model is available online[13].

the *preservation policy*. Finally, user stories which refer to topics covered by a preservation policy also refer to *storage* types and *volume* of data.

The visualisation allows us to understand which DMP themes are closely connected, who the stakeholders are that need this information and in which phases this information is processed. This matches with the goals of our consultation.

4.2.4 *From user stories to requirements and models*

The visualisation of user stories allowed us to see high level relations between stakeholders, phases and DMP themes. It also allowed us to see which information must be included in the model in the first place and which is a nice to have addition. For example, *metadata* was found relevant to all stakeholders in all phases and is related with most of the themes, while description of *IT resources* needed to run a project was only referred by *institutional* stakeholder in the *planning* phase and was only connected to *storage*.

By performing the stakeholder consultation we also aimed to define requirements for a common data model, that is, identify which specific information must be included in the model. For this reason we again used GitHub to filter out user stories for each *yellow label* – that is for each label referring to scope of information covered by a DMP. Then in a text document we wrote down specific requirements expressed by the user stories falling under the selected label. For each requirement we noted a number of a user story from which we derived the requirement. For example, for the *reuse* label we derived following requirements:

- Reuse
  - Links to (meta-)data location [89, 90, 56, 39, 60]
  - IPR (can be reused or not) [11, 69, 41, 30, 53]
  - Privacy [29]
  - Link to 'DMP Corresponding author' [9, 88]

– Links to related sources e.g. website, documentation [10, 25]

– Deployment scenario [16]

The full list of requirements can be found online[14].

We analysed the requirements and noticed that similar requirements exist in different context. For example, information about the 'DMP corresponding author' is relevant not only for the *reuse*, but also for the *administrative* information. For this reason, we re-arranged the document and grouped the requirements into five major categories:

1. Administrative, Roles and Responsibilities[15]
2. Data[16]
3. Infrastructure[17]
4. Security, Privacy and Access Control[18]
5. Policies, legal and ethical aspects[19]

Each category contains requirements that we derived from analysing user stories after re-arranging and grouping. They constitute an initial set of requirements for common model for machine-actionable DMPs. Due to space limitation we do not provide a full list in the paper. They can be found online (links are provided in the item list above for each category)

5. RESULTS AND DISCUSSION

Of the 108 users stories collected, 77 were accepted as being directly applicable to building a data model for maDMP. Another 22 user stories were assessed as being out of scope and an additional 9 were rejected completely (see Figure 6).

While providing valuable insights into the maDMP process, most of the out of scope stories could not be directly translated into specific requirements for a DMP data model. The DMP Common Standards working group is mindful of scope creep and the need to devote limited resources to the task of creating a data model, so anything that was not directly related to this has been deemed out of scope. Many out of scope stories relate to how the ecosystem of services (cf. Section 3.2) would actually operate, or how workflows would integrate into a DMP system - all very worthwhile and helpful in understand expectations of stakeholders towards maDMPs, but not directly relevant to specific task of creating a model.

We are planning to revisit out of scope user stories at a later stage once a model has been developed. Then these use stories can become useful in formulating use cases and pilots in which the model is tested. The out of scope user stories are considered as relevant by the RDA Exposing DMPs working group that defines new use cases in which information from DMPs can be reused in different contexts[20].

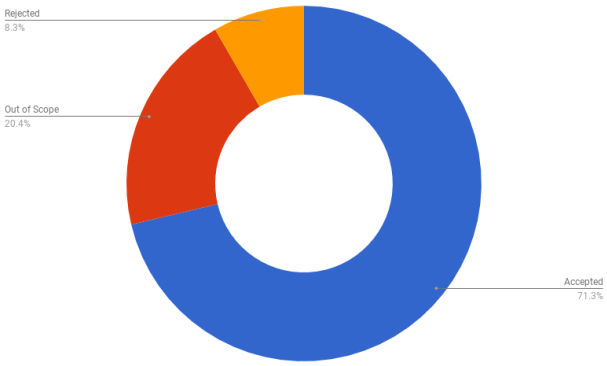

Figure 6: User stories accepted, rejected or out of scope

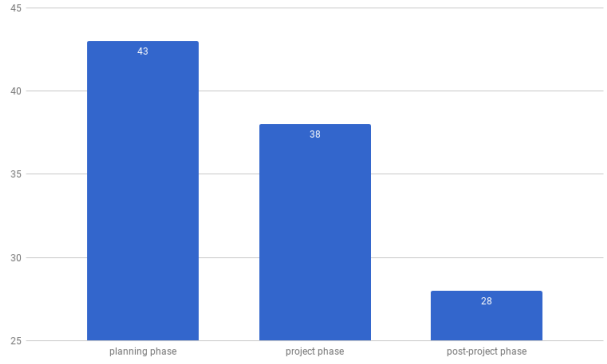

Figure 7: User stories with label for phase

The rejected issues could not be easily translated into requirements as they were either too vague or not related directly to maDMPs. Again, these issues have been retained and labelled and will be revisited at a later stage.

As expected the phase with the most user stories (43) was the *planning* phase (see Figure 7) and we would expect data management to be highly associated with the *planning* phase of a research project. However there are still a considerable number of user stories that were related to the active *project* phase (38) and the *post-project* phase (28) indicating that data management planning occurs throughout the entire research lifecycle.

A range of stakeholders were associated with user stories (Figure 8) with *repository operators* being the most common (even outnumbering the researcher). This suggest that repository systems could be a good first candidate for implementing machine actionable processes.

Many different subjects were associated with the user stories (Figure 9). The subject *reuse* was the highest, with *metadata* a close second. The third most popular subject

[14] https://docs.google.com/document/d/1sWVy0Rqj9fGsjs6GyFnBd3fH6XF2088zjK8U-1wLq4c/edit?usp=sharing

[15] https://drive.google.com/open?id=1rEcX_9zf3vuoj0ap8UBE4xhQa3B4OvxbE-rY92cVcdA

[16] https://drive.google.com/open?id=1GRBxgOKf5VGfJ9YGzcQqID2qn6V5PKcwNUlAYGsJwj0

[17] https://drive.google.com/open?id=1L800iw89Pqh-sx4-HmmqNtKhV7GaZ8ANshhYEu_wjM4

[18] https://drive.google.com/open?id=1D1grHL9TQsvtOoD6Oos5lXW4QG9WMf1xA5zgZ8tL3VI

[19] https://drive.google.com/open?id=1RhsT7JYVcYJta6S040s2QV99iz1ePQOhSzsbrLxAous

[20] https://www.rd-alliance.org/groups/exposing-data-management-plans-wg

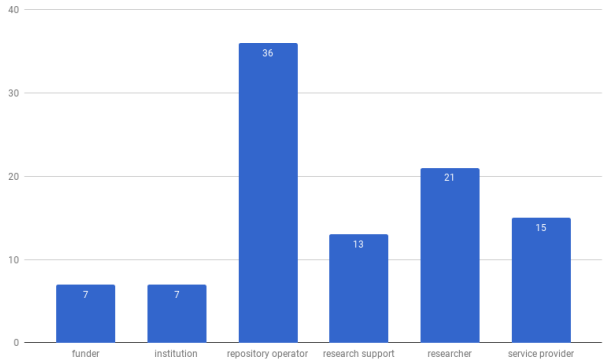

Figure 8: User story stakeholders

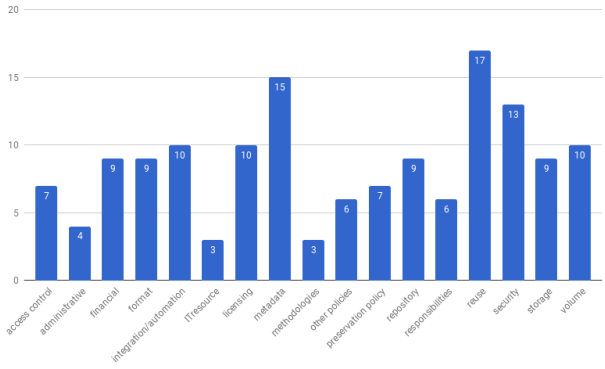

Figure 9: User story subjects

label was *security*.

All subjects were present in at least 3 user stories and most were present in 6 or more, so any data model produced will need to accommodate the breadth of subjects represented here. It is anticipated that there is a need for a data model that can include extensions to accommodate all the different subjects across different domains.

While this process has provided valuable insights into the stakeholders, phases and themes associated with DMPs, there is still significant information that is incomplete. In many cases the granularity of the information is too high for meaningful analysis. For example, we know that file size is important, but we do not know if an estimation is sufficient, or if we need the exact size. We also know which stakeholders are involved and the phase and subjects they are associated with, but we do not know what kinds of information stakeholders need and how those needs might change. Stakeholders may require different information depending on the phase of the lifecycle or for different subjects. For this reason, we are currently undertaking a second consultation with experts in each field to discuss specific fields and standards currently in use. This is also because our first consultation focused on broad analysis of requirements towards maDMPs - understanding their scope and involved stakeholders. The aim of the second one is to extend the list of identified requirements by diving into details.

6. CONCLUSIONS

In this paper we presented differences between traditional and machine-actionable DMPs by comparing how each of them models information. We also described an ecosystem of services acting on behalf of human stakeholders involved in research data management.

Furthermore, we applied requirements engineering methodology known in software and system engineering to derive requirements for a common data model for machine-actionable DMPs. We used GitHub as a platform for collating and processing user stories. The approach was successful and resulted in a significant collection of data that was further developed into requirements for the common data model for machine-actionable DMPs.

In many ways we are building on the work of DMP tool developers who have created tools and systems that reflect the viewpoint of specific stakeholders (primarily funders and researchers). However, the consultation described in this paper is the first attempt that we know of to use a bottom-up approach to elicit the complex interrelations between all DMP stakeholders. Thus, our research contributes to removing ambiguities among stakeholders and establishing a common notion of machine-actionable DMPs.

The next phase of the project is a deeper dive to develop specific requirements by utilizing experts in particular fields. The labelling and categorization of these users stories has assisted with identifying the most promising areas that can be targeted in this next phase.

7. ACKNOWLEDGMENTS

The stakeholder consultation would not be possible without support of enthusiastic RDA DMP Common Standards group members. 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).

8. REFERENCES

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

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

[3] EUDAT. License selector tool. https://eudat.eu/services/userdoc/license-selector. Accessed: 2018-02-01.

[4] European Commission. EOSC Declaration. http://ec.europa.eu/research/openscience/pdf/eosc_declaration.p 2017. Online; accessed 29 March 2018.

[5] S. Higgins. The DCC Curation Lifecycle Model. *International Journal of Digital Curation*, 12 2008.

[6] U. B. Kejser, J. Davidson, D. Wang, S. Strodl, T. Miksa, K. H. E. Johansen, A. B. Nielsen, and A. Thirifays. State of the art of cost and benefit models for digital curation, June 2014.

[7] W. K. Michener. Ten simple rules for creating a good data management plan. *PLoS Comput Biol*, 11,

10/2015 2015.

[8] T. Miksa and K. Ashley. Workshop report - Research Data Lifecycle and Machine-actionable Data Management Plans, 2017.

[9] T. Miksa, A. Rauber, and R. Vieira. Vplan - ontology for collection of process verification data. In *Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014, Melbourne, Australia, October 6 - 10, 2014*, 2014.

[10] S. Simms, S. Jones, D. Mietchen, and T. Miksa. Machine-actionable data management plans (madmps). *Research Ideas and Outcomes*, 3:e13086, 2017.

[11] Tomasz Miksa and Paul Walk and Peter Neish. RDA WG DMP Common Standards Case Statement. https://www.rd-alliance.org/group/dmp-common-standards-wg/case-statement/rda-wg-dmp-common-standards-case-statement, 2017. Online; accessed 29 March 2018.

[12] 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. *Nature Scientific data*, 3, 2016.