Exploring Machine-Actionable Data Management Plans

João Manuel Fernandes Cardoso[1][0000−0003−0057−8788]

Instituto Superior Técnico
Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
joao.m.f.cardoso@tecnico.ulisboa.pt

Abstract. The product of scientific research is the production of data. Nowadays researchers are faced with the growing challenge of how to manage, preserve and publish large sets of data, in way that allows for it to be reproduced and reused. This paper proposes to explore the concept of machine-actionable data management plan. In particular, the usage of semantic techniques to both express and exploit the features of machine-actionable data management plans will be analysed. As semantic techniques have been proven to be a reliable means to represent and analyse data in other contexts.

Keywords: Machine-Actionable Data Management Plan · Research Data · Ontologies.

1 Introduction

One of the products of scientific research is the production of data. Either as a direct product of research, or as raw information to be processed and interpreted for knowledge production. Nonetheless the exercise of performing scientific research results in ever increasing quantities of data [5]. Moreover, there is little incentive for researchers to publish their data, regardless of the success of their endeavours [8], thus limiting the reproducibility transparency of methods and workflows. Researchers are therefore faced with the growing challenge of how to manage, preserve and publish their data, in way that allows for it to be reproduced and reused.

Research data management (RDM) [14,12] is one of the approaches to tackle these challenges. It is a manifestation of the 'open science' concept, in which scientific research and data is to be made freely accessible [10,17]. The data management plan (DMP) [6], is one of the tools available for RDM. It is a document describing the techniques, methods and policies on how data from a research project is to be created, documented, accessed, preserved and disseminated. The machine-actionable DMP (maDMP) [11] (sometimes referred as "active, "dynamic", or "machine-readable" DMP) reflects an attempt to build upon the concept of DMP, by fitting it with dynamic features.

The objective of this research is to explore the concept of maDMP. Specifically, the usage of semantic techniques to both express and exploit the features of maDMP documents will be analysed.

In the field of computer science the most commonly accepted definition of the term "ontology" describes ontologies as a "formal, explicit specification of a shared conceptualisation" [13]. Breaking down the definition, and putting it into context, we can define that: "conceptualisation" could refer to an abstract model of an maDMP, "explicit" means that the type of concepts used and the constraints on their use are to be explicitly defined, "formal" refers to the fact that the ontology should be machine readable, and "shared" reflects that ontology should capture consensual knowledge accepted by the community [13].

1.1 Research Questions

To achieve the proposed objective, this work considers the following assumption, to then allow for the exploration of the related research questions:

Assumption:

- Existing state of the art on semantic technologies, and in particular ontology representation and analysis[1,13], raises the possibility that these technologies can also be applied to exploit the features of the maDMP through its semantic representation.

Research Questions:

- **Formal representation of a maDMP: How can a maDMP be formalised?** - The question focuses on the possibility of developing or taking advantage of any existing ontologies to express an maDMP, and add structure to the knowledge it contains.
- **maDMP services: Can a maDMP model, represented as an ontology, be used to develop services that use information contained within the model to produce value?** - The objective of this question lies in the validity of creating services that take advantage of semantic techniques to produce value or act upon knowledge expressed in the maDMP.

The rest of this paper follows the following outline: Section 2 presents an overview on the fundamental concepts associated with RDM, DMP and maDMP. Finally in section 3 the proposed research is described. Particular focus is given to the research context, in which the validation scenarios are presented , and the work plan.

2 Research Data Management Fundamentals

2.1 Research Data Management

RDM 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 possible 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" [16]. RDM is best perceived when illustrated through the data life cycle [14], as seen in figure 1.

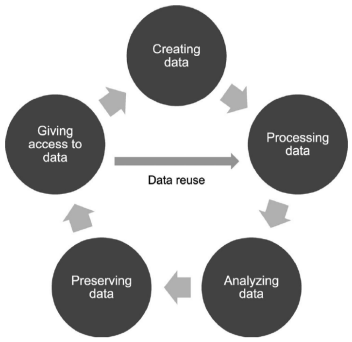

Fig. 1. Data life cycle [14]

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.

2.2 Data Management Plan

Engaging in scientific research often leads to the production of large sets of data. With such large quantities of data, it is imperative that any RDM practices that are applied, be clearly documented and accessible. The concept of DMP was introduced as a possible approach to cover that necessity. 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. Presently various funding bodies require that any funding application be accompanied by a DMP.

Principles and Practices According to literature [6], 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 organized

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

2.3 Machine-Actionable Data Management Plan

DMPs are meant to be revisited and updated throughout the project data life cycle. However, in practice the DMP documents are created either at the onset or close of a project and are rarely updated. This results into a mostly static document. The level of detail of a DMP also varies according to the expertise and meticulousness of its creators. The lack of consistency may lead to the DMP not having the necessary information to detail its principles and practices, or offering broad or unclear descriptions. Additionally there is no standard set of requirements for a DMP. Each funding body establishes its own requirements, and these can vary both in depth and breadth of detail. This means that before creating a DMP, researchers should take the time to understand the requirements set by the funding body, either by analysing the online proposal guides or any existing public request for proposals (RFP). Even though there are publicly available DMP templates and examples (e.g. DMPTool[1] or DMPonline[2]), there is a considerable overhead to the creation of a DMP. Moreover, any DMP created based on these existing templates, is likely to be nothing more than an unstructured collection of knowledge. All of these points only strengthen the fact that, although ideally a DMP should be dynamic in nature and a fundamental part of data management in a project. In reality, it is often considered nothing more than bureaucratic hassle.

The concept of maDMP [11] (sometimes referred as "active, "dynamic", or "machine-readable" DMP) was born from an initiative to introduce dynamic features to a DMP, and thus realise its true potential. This concept was brought to light, in conferences such as the International Digital Curation Conference 2017 (IDCC17)[3], or through the recently created Research Data Alliance (RDA)[4] work groups on Active DMP[5] and DMP Common Standards[6], it would benefit from being enriched with machine-readable added value.

maDMP Principles and Practices As is the case for the DMP (see section 2.2), there is also a proposal for a set of 10 principles and practices that are

[1] https://dmptool.org/
[2] https://dmponline.dcc.ac.uk/
[3] http://www.dcc.ac.uk/drupal/events/idcc17
[4] https://www.rd-alliance.org
[5] https://www.rd-alliance.org/groups/active-data-management-plans.html
[6] https://www.rd-alliance.org/groups/dmp-common-standards-wg

specific to the maDMP [7]. 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 DMP documents 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 maDMP documents
7. Make maDMP documents available for human and machine consumption
8. Support data management evaluation and monitoring
9. Make DMP documents updatable, living, versioned documents
10. Make DMP documents publicly available

3 Proposed Research

Training is key in the preparation of researches to perform data management tasks, for RDM tasks can vary according to the field of science, the project size and the type of data [3]. These two points make the adoption of the 'open science' concept difficult, for without standardised data annotation and documentation it is pointless to focus on preservation and sharing. One of the attempts to mitigate these points led to the introduction of the DMP [15,2]. The more common form of a DMP is a document that aids researchers in describing how their research project data is to be managed throughout its life cycle. Funding bodies, in recognition that having a plan before starting a research project is a valuable first step, have started to require that any funding applications be accompanied by a DMP. This DMP is expected to be revisited and updated throughout the project data life cycle. Unfortunately this is not always the case. A DMP is often created either at the start or end of a research project and are then rarely revisited. As a result DMP documents are perceived as a mostly static document. In order to counteract against this established perception, and provide the DMP with dynamic features that the concept of maDMP [11] was introduced. With an maDMP any described information can be acted upon not only by the researchers but also by machines. This opens a wide range of possible applications that were unavailable with a standard DMP.

The objective of this research is therefore to explore the features made available by the maDMP. Specifically, it is raised the hypothesis that the usage of semantic techniques to both express and exploit the features of maDMP documents would be of value. The following are possible scenarios of application where an maDMP expressed using semantic technologies could be of value: In a scenario where a researcher was trying to publish a given dataset. Any conflicting policies in the maDMP could be automatically retrieved. Retrieved policies

could then be acted upon either by the researcher or other services, thus allowing the dataset to be published. Another scenario could be automated repository selection. Researchers looking for repositories to store their data could be advised on various compliant repositories, based on the nature of their data and other policies within the maDMP.

3.1 Research Context

To prove the validity of the proposed approach, the following use cases will be to be considered (others might arise during the research): The RDA DMP Common Standards Working Group aims at developing a common information model, and specifying access mechanisms that make DMP documents machine-actionable; ELIXIR [4] is an organisation that supports the bioinformatics community across Europe, allowing researchers to find and share, data and knowledge; PRECISE [9] is a joint research program that aims to explore the application of the concept of prevention and treatment strategies that take individual variability into account. As a result, large amounts of research data are generated.

Regarding the first use case, the research work will be performed in collaboration with the working group. The objective is to aid in the attainment of the working group's objectives, in particular in the development of a common information model for the maDMP. The last two use cases have in common the need to manage research data whose domain might prove to be a challenge. This is not due to its large quantities, but to the fact that it might be under strict privacy policies, or other constraints.

The overall objective of exploring the usage of the maDMP and any associated services in both use cases, is to facilitate information exchange, management and embed maDMP documents in the existing workflows.

3.2 Work Plan

To attain the objective of this work, the following three steps are proposed to be taken: Study the common structure of an maDMP. With that information, it might then be possible to create a maDMP model that represents specific type of scenario (e.g. A dataset that registers data on stroke patients); Express the created model in a representation language that allows for machine-readability, and therefore add value to the data contained within the maDMP; Explore the creation and implementation of services that that advantage of the features of the maDMP to automate the management, preservation and reuse of data.

4 Acknowledgements

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).

References

1. Breitman, K., Casanova, M.A., Truszkowski, W.: Semantic web: concepts, technologies and applications. Springer Science & Business Media (2007)
2. Corrall, S., Kennan, M.A., Afzal, W.: Bibliometrics and research data management services: Emerging trends in library support for research. Library trends **61**(3), 636–674 (2013)
3. Editorial: Everyone needs a data-management plan. Nature **555**(7696), 286 (2018). https://doi.org/10.1038/d41586-018-03065-z
4. ELIXIR Hub: ELIXIR Europe, `https://www.elixir-europe.org`
5. Gray, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A., DeWitt, D.J., Heber, G.: Scientific data management in the coming decade. Acm Sigmod Record **34**(4), 34–41 (2005)
6. Michener, W.K.: Ten simple rules for creating a good data management plan. PLoS computational biology **11**(10), e1004525 (2015)
7. Miksa, T., Simms, S., Mietchen, D., Jones, S.: Ten simple rules for machine-actionable data management plans (preprint) (Feb 2018). https://doi.org/10.5281/zenodo.1172673, `https://doi.org/10.5281/zenodo.1172673`
8. Nosek, B.A., Spies, J.R., Motyl, M.: Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science **7**(6), 615–631 (2012)
9. Precise Program: Precise Program, `https://vre.precisemed.org/`
10. Schiermeier, Q.: Data management made simple. Nature **555**(7696), 403–405 (2018). https://doi.org/10.1038/d41586-018-03071-1
11. Simms, S., Jones, S., Mietchen, D., Miksa, T.: Machine-actionable data management plans (madmps). Research Ideas and Outcomes **3**, e13086 (2017). https://doi.org/10.3897/rio.3.e13086, `https://doi.org/10.3897/rio.3.e13086`
12. Strasser, C.: Research data management. National Information Standards Organization (2015)
13. Studer, R., Benjamins, V.R., Fensel, D., et al.: Knowledge engineering: principles and methods. Data and knowledge engineering **25**(1), 161–198 (1998)
14. Surkis, A., Read, K.: Research data management. Journal of the Medical Library Association: JMLA **103**(3), 154 (2015)
15. Tenopir, C., Sandusky, R.J., Allard, S., Birch, B.: Research data management services in academic research libraries and perceptions of librarians. Library & Information Science Research **36**(2), 84–90 (2014)
16. Whyte, A., Tedds, J.: Making the case for research data management. Digital Curation Centre (2011)
17. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3** (2016)