# Machine-Actionable Data Management Plans: A Knowledge Retrieval Approach to Automate the Assessment of Funders' Requirements

João Cardoso[1,2(✉)], Diogo Proença[1,2], and José Borbinha[1,2]

[1] INESC-ID, Lisbon, Portugal
[2] Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
{joao.m.f.cardoso,diogo.proenca,jlb}@tecnico.ulisboa.pt

**Abstract.** Funding bodies and other policy-makers are increasingly more concerned with Research Data Management (RDM). The Data Management Plan (DMP) is one of the tools available to perform RDM tasks, however it is not a perfect concept. The Machine-Actionable Data Management Plan (maDMP) is a concept that aims to make the DMP interoperable, automated and increasingly standardised. In this paper we showcase that through the usage of semantic technologies, it is possible to both express and exploit the features of the maDMP. In particular, we focus on showing how a maDMP formalised as an ontology can be used automate the assessment of a funder's requirements for a given organisation.

**Keywords:** Data Management Plan · Machine Actionable Data Management Plan · Semantic technologies

## 1 Introduction

Funding bodies and other policy-makers are increasingly more concerned with Research Data Management (RDM). One of the contributing factors is the general perception that research data should be a public good [16]. In order to guide researchers through the process of managing their data, many funding agencies (e.g. the National Science Foundation (NSF), the European Commission (EC), or the Fundação para a Ciência e Tecnologia (FCT) have created and published their own open access policies, as well as requiring that any grant proposals be accompanied by a Data Management Plan (DMP).

The DMP [7] is one of the tools available to researchers to aid in the management of their data. The DMP is a document describing the techniques, methods and policies on how data from a research project is to be created or collected, documented, accessed, preserved and disseminated. The DMP is not without issues, such as lack of standardisation, lack of continuous updates throught the


J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 118–125, 2020.
https://doi.org/10.1007/978-3-030-45442-5_15

project life cycle, etc. As a result of these issues, the DMP is seen more as bureaucratic obligation, than as a valuable asset for data management.

The concept of Machine-Actionable DMP (maDMP) [13] (sometimes referred as "active", "dynamic", or "machine-readable" DMP), aims at addressing some of these issues by making the DMP machine-readable without compromising its human-readability. The adoption of an open, shared and interoperable concept of maDMP could bring multiple benefits, such as facilitating data discovery and reuse, and enabling automated evaluation and monitoring, etc.

It is clear that having funding agencies and universities push for DMP usage is not enough to achieve true adoption and standardisation. Researchers need to be convinced that the DMP can be a major tool to support RDM, especially if it is perceived as a living object, and there has to be a move towards standardising DMP documents. One of the attempts to tackle the standardisation issue, is being carried out by the Research Data Alliance (RDA) DMP Common Standards Working group [8]. Whose objective was to establish a common data model that would define a core set of elements for a DMP.

The objective of this paper is to explore the application of semantic technology to RDM. In particular, by resorting to semantic technologies to both express and exploit the features of the maDMP [4,14]. To that effect we set out to show a DMP formalised as an ontology can be used automate the assessment of a funder's requirements for a given organisation.

This paper is organised as follows. Section 2 offers definitions on the concepts of DMP, maDMP and Semantic Technologies. Section 3 describes our approach on how to establish a DMP creation service that allows for semantic based DMP representation. Section 4 closes the paper by presenting both a final appreciation on the proposed approach, and a description of possible future work on this topic.

## 2 Related Work

**Ontologies.** Semantic Technology has shifted from originally tackling syntax and structural issues, to focus on the exploitation of the semantics of information [11] to promote both system interoperability and enhance the web infrastructure [12].

Ontologies play a key role in Semantic Technology, for they enable knowledge representation through formal semantics. Studer describes ontologies as a "formal, explicit specification of a shared conceptualization" [14]. Ontology usage can be sorted into three categories [15]: Human communication by providing a common interpretation of knowledge, interoperability by enabling data exchange among heterogeneous sources, and systems engineering by providing a shared understating of problems.

Ontologies can be represented by formal languages, that have the dual purpose of encoding knowledge on specific domains, as well as including reasoning rules that aid in the processing of that knowledge. These formal languages are referred to as ontology languages (e.g. Resource Description Framework (RDF), Web Ontology Language (OWL), etc.).

**Semantic Reasoning.** According to Lenzerini et al. [5], "several reasoning tasks can be carried out to deduce implicit knowledge from the explicitly represented knowledge". Hence, reasoning mechanisms are used for carrying two tasks, ontology validation and ontology analysis. The validation of an ontology consists in checking if the ontology is correctly modelling the domain in focus, whereas the analysis of an ontology focuses on deducing facts about the modelled domain, processing and extracting new information from the original information. The same author also proposes a classification for different types of reasoning according to the results aimed for: deduction, induction and abduction.

Reasoners, sometimes referred as Discription Logics (DL) systems, can be organized into three generations [2]. The third and last generation of reasoners focus on optimized reasoners which are expressive, sound and complete. In this last generation are included FaCT++, RacerPro and HermiT.

**Semantic Based Assessment.** The relevance of ontology-based techniques is demonstrated by the growing use of ontologies in a diversity of domains.

Antunes proposed a "model for the representation and integration of ontologies" [1] which allows the analysis of an architecture. The model is based on meta-model and model integration consisting of an upper ontology and a collection of domain specific ontologies.

Bakhshandeh [3] proposed model for the representation, integration and analysis of EA models. The proposed model addresses the need for a representation that allows integration of different metamodels and models along with their analysis by computational needs. This proposal is based on the hypothesis that ontologies can represent, integrate and support the analysis of enterprise architecture models.

Proença [10] proposed a model and method to represent maturity models, namely its components, rules and assessment criteria using ontologies. It demonstrated how to use computational inference as an analysis technique to take advantage of the information already encoded in maturity models with the purpose of automating existing maturity assessment methods.

## 3 Proposed Approach and Validation

**Method.** This section details how the objective of this paper was approached, as well as a description of the methods used to provide a preliminary validation of said approach. Additionally, examples of the potential that semantic technologies can provide for maDMP exploitation are also given.

The overarching goal of this work to both express and exploit the features of the maDMP by resorting to semantic technologies. In particular, it Pfocuses on showing how a DMP formalised as an ontology can be used to automate the assessment of a funder's requirements for a given organisation.

Our approach for maDMP generation comprises of 8 tasks, that can be analysed in Fig. 1, can be interpreted into three main parts. The first part comprises of the execution of the first two tasks, which resulted into the creation of an

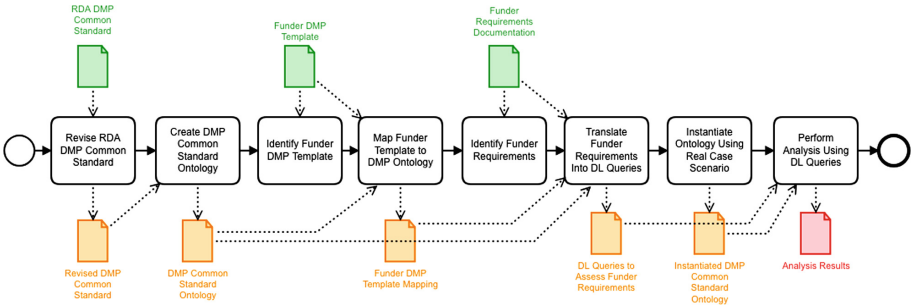


**Fig. 1.** The process for maDMP generation

ontology compliant with the DMP Common Standards model, that is detailed in this section. The second part comprises of the execution of the following four tasks and results in both the collection of the necessary mappings between the ontology and the identified DMP templates, and creation of DL queries based on the funders' requirements. The final part comprises of the last two tasks, and results in the creation of a poExpulated ontology that is subsequently analysed using the previously created DL queries. The second and third part of the approach are detailed in the last subsection.

With our approach researchers should be able to select a template and fill the necessary forms to generate a DMP. Our focus however is on having the generated DMP be machine-actionable, DMP Common Standards Model compliant, and expressed through the usage of semantic technologies.

**DMP Common Standard Ontology.** The DMP Common Standards Working group [8] was created with the objective of establishing a common data model that would define a core set of elements for a DMP. The resulting data model has a modular design allowing it to be extended by existing standards and vocabularies. The DMP Common Standards working group is also meant to provide reference implementations of the model in popular formats (e.g. JSON, XML, RDF).

Our approach called for the revision of the DMP Common Standards Model, and this resulted on a representation of the model using semantic technologies. The DMP Common Standard Ontology (DCSO)[1], was created with the objective of providing an implementation of the DMP Common Standards model expressed through the usage of semantic technology, which has been considered a possible solution in the data management and preservation domains [9]. The DCSO is represented using OWL [6], and its model can be analysed in Fig. 2.

The following step was to collect DMP templates from funding agencies. To that effect we collected the DMP template for both the EC Horizon 2020

[1] All the ontologies mentioned in this paper can be found in the DMP Common Standard Ontology Repository: https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/tree/master/ontologies.

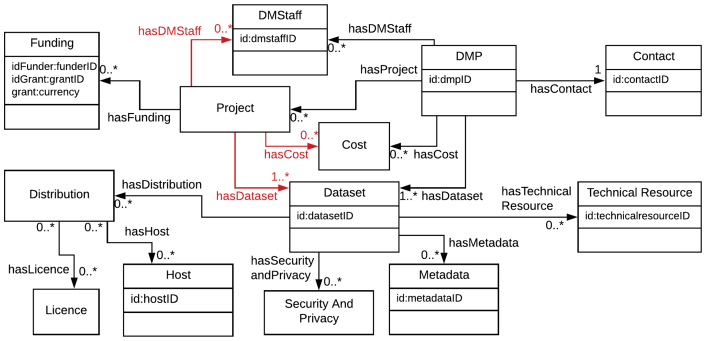


**Fig. 2.** The DMP common standards ontology with the proposed extension highlighted (Color figure online)

programme, and the FCT project funding. We then proceeded to attempt to validate that DCSO would cover the entirety of both the DMP templates, and establish the necessary mappings. However we came to the conclusion that we would have to extend the DCSO, in order to better address the mapping of the DMP templates to the DCSO. The extension was limited to the addition of three object properties that were added to the Project class, they can be identified in Fig. 2 highlighted in red.

**Funders' Requirements Assessment Using DL Queries.** Given the mappings between the DPM templates and the DCSO, the next step was to collect the funder's requirements for a DMP, and have them translated into DL Queries that were executable over the DCSO. It was however necessary to first instantiate the extended version of the DCSO. To that effect we resorted to areal case scenario, the Genomics Unit at the Instituto Gulbenkian de Ciência[2].

The Genomics unit provides Next Generation Sequencing services using state-of-the-art Illumina sequencers, and is currently engaged in two projects, namely *Oneida*[3] and *GenomePT*[4]. Our decision was to create a single DMP that would cover both projects. By using the instantiated ontology and the DL queries we were able to assert if the DMP complied with funder's requirements and compiled the analysis results in a document.

Figure 3 shows the verification of one of the FCT requirements, which state that every project must cover the cost to provide open access to their outputs. Some journals charge an additional fee in order to provide open access to a published paper, as a result researchers must include the costs of these fees in the DMP to be submitted in the project proposal, as well as, in the periodic reviews. As can be seen in Fig. 3 the costs for open access are detailed for the *GenomePT* project bot not for the *Oneida* project. Additionally, we can use these queries to get the list of costs covered by a project or the whole DMP.

[2] http://facilities.igc.gulbenkian.pt/genomics/genomics.php.
[3] https://www.itqb.unl.pt/oneida.
[4] https://www.genomept.pt/.

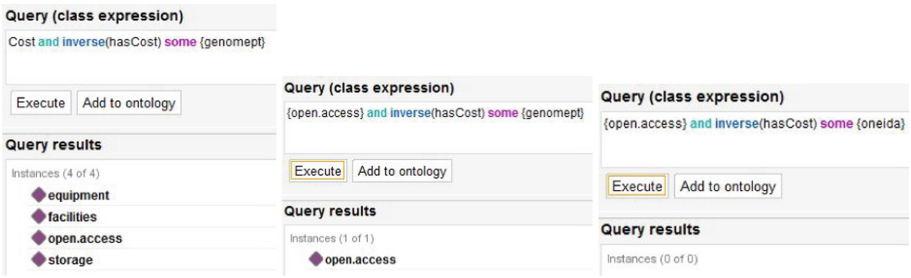


**Fig. 3.** Verifying if the *GenomePT* and *Oneida* projects cover costs related with open access.

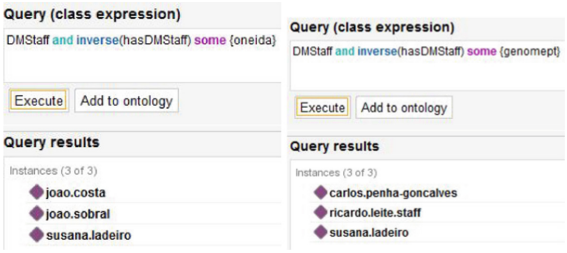


**Fig. 4.** Verifying the researchers assigned to *GenomePT* and *Oneida* projects.

Another example is the researchers effort per project. Figure 4 shown the staff assigned to each project. The researcher "susana.ladeiro" is assigned to both projects which means that this researcher does not work in just one project. This means that the costs associated with this researcher must be covered in both projects.

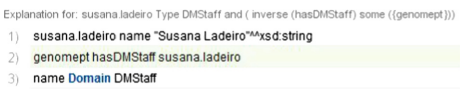


**Fig. 5.** Explanation for the results provided by a DL query.

Another functionality provided by the use of ontologies is the provision of explanations for each of the results of a query. In this example, Fig. 5 depicts the explanation for assigning the researcher "susana.ladeiro" to the *GenomePT* project. This is a very simple example to show the potential of this feature, it can provide researchers an explanation on which funder's requirements are not compliant and the reasons for being non-compliant.

## 4 Conclusions and Future Work

The overall goal of this paper was to demonstrate the use of semantic technologies to both express an maDMP and exploit its features. With the approach described in Sect. 3, we focused on showcasing how a DMP expressed as an ontology can impact the assessment of funder's requirements for two organisations.

This paper is a report on an ongoing effort. As such, there are still action points that we consider in need of further development. The DCSO, due to the nature, is often too generic to adapt to specific DMP templates. That implies that more DCSO extensions will have to be created to cater for specific contexts. Another possible action point is use one of the many existing DMP creation frameworks (e.g. DMP Online[5], DMP Tool[6], Data Stewardship Wizzard[7] etc.) to automate the creation a instantiated DCSO. This would minimize the necessity for an ontology expert in the process.

**Acknowledgements.** This work was supported by national funds through FCT with reference UID/CEC/50021/2019, and by project PRECISE (LISBOA-01-0145-FEDER-016394).

## References

1. Antunes, G.: Analysis of enterprise architecture models: an application of ontologies to the enterprise architecture domain. Ph.D. thesis, Instituto Superior Técnico, Universidade de Lisboa (2015)
2. Baader, F., Calvanese, D., McGuinness, D., Patel-Schneider, P., Nardi, D.: The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, Cambridge (2003)
3. Bakhshandeh, M.: Ontology-driven analysis of enterprise architecture models. Ph.D. thesis, Instituto Superior Técnico, Universidade de Lisboa (2016)
4. Breitman, K., Casanova, M.A., Truszkowski, W.: Semantic Web: Concepts, Technologies and Applications. Springer, Heidelberg (2007). https://doi.org/10.1007/978-1-84628-710-7
5. Lenzerini, M., Milano, D., Poggi, A.: Ontology representation & reasoning. Technical report, NoE InterOp (IST-508011) (2004)
6. McGuinness, D.L., Van Harmelen, F., et al.: Owl web ontology language overview. W3C Recommendation **10**(10), 2004 (2004)
7. Michener, W.K.: Ten simple rules for creating a good data management plan. PLoS Comput. Biol. **11**(10), e1004525 (2015)
8. Miksa, T., Neish, P., Walk, P.: WG DMP common standards case statement (2017)
9. Miksa, T., Vieira, R.J.C., Barateiro, J., Rauber, A.: VPlan-ontology for collection of process verification data (2014)
10. Proença, D.: Maturity assessment support with conceptual modelling methods and semantic techniques. Ph.D. thesis, Instituto Superior Técnico, Universidade de Lisboa (2018)

[5] http://dmponline.dcc.ac.uk.
[6] https://DMPTool.org.
[7] https://ds-wizard.org.

11. Sheth, A.P.: Changing focus on interoperability in information systems: from system, syntax, structure to semantics. In: Goodchild, M., Egenhofer, M., Fegeas, R., Kottman, C. (eds.) Interoperating Geographic Information Systems. The Springer International Series in Engineering and Computer Science, vol. 495, pp. 5–29. Springer, Boston (1999). https://doi.org/10.1007/978-1-4615-5189-8_2
12. Sheth, A.P., Ramakrishnan, C.: Semantic (web) technology in action: ontology driven information systems for search, integration, and analysis. IEEE Data Eng. Bull. **26**(4), 40 (2003)
13. Simms, S., Jones, S., Mietchen, D., Miksa, T.: Machine-actionable data management plans (maDMPs). Res. Ideas Outcomes **3**, e13086 (2017)
14. Studer, R., Benjamins, V.R., Fensel, D., et al.: Knowledge engineering: principles and methods. Data Knowl. Eng. **25**(1), 161–198 (1998)
15. Uschold, M., Gruninger, M.: Ontologies: principles, methods and applications. knowl. Eng. Rev. **11**(02), 93–136 (1996)
16. Whyte, A., Tedds, J.: Making the case for research data management. DCC Briefing Papers (2011)