IJDC | General Article

Information Integration for Machine Actionable Data Management Plans

Tomasz Miksa
TU Wien, SBA Research

Andreas Rauber
TU Wien

Roman Ganguly
University of Vienna

Paolo Budroni
University of Vienna

Abstract

Data management plans are free-form text documents describing the data used and produced in scientific experiments. The complexity of data-driven experiments requires precise descriptions of tools and datasets used in computations to enable their reproducibility and reuse. Data management plans fall short of these requirements. In this paper, we propose machine-actionable data management plans that cover the same themes as standard data management plans, but particular sections are filled with information obtained from existing tools. We present mapping of tools from the domains of digital preservation, reproducible research, open science, and data repositories to data management plan sections. Thus, we identify the requirements for a good solution and identify its limitations. We also propose a machine-actionable data model that enables information integration. The model uses ontologies and is based on existing standards.

Received 28 November 2016 ~ *Accepted* 21 February 2017

Correspondence should be addressed to Tomasz Miksa, Favoritenstraße 16, 1040 Wien. Email: tmiksa@sba-research.org

An earlier version of this paper was presented at the 12th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v12i1.529
DOI: 10.2218/ijdc.v12i1.529

Introduction

Data Management Plans (DMPs) are a compulsory document accompanying project proposals or project deliverables submitted to research funders or further funding bodies. They describe which data is used and produced in an experiment, where the data is archived, which licenses and constraints apply, and to whom credit should be given. DMPs were introduced to make data Findable, Accessible, Interoperable, and Reusable (FAIR) (Wilkinson et al., 2016). DMPs are created manually by researchers who use checklists. There are also online tools that provide questionnaires tailored to the specific funder's needs[1,2].

Despite the efforts of the digital curation community to familiarise researchers with the best practices in data management, the information provided in DMPs is vague. The quality of information not only depends on the expertise and scrupulousness of the person writing the DMP, but also on the stage at which the DMP is created. When a DMP is written at the end of a project, important data may not be available anymore. A DMP is supposed to be a living document, but it is sometimes considered as a redundant bureaucracy. It is created and updated not when the data is actually produced, but when it is required for reporting. Due to its free-form text, questions can remain unanswered, or the answers can be very generic.

This problem has been recognised by the community and is being discussed within the Research Data Alliance. Participants of the CERN workshop discussed shortcomings of DMPs and ways of addressing them. They identified 'encodings for exporting DMPs' as one of the next developments needed[3]. Automation and machine-actionability were identified as key factors enabling deployment of European Open Science Cloud (European Commission, 2016). In parallel, a wide range of tools and concepts were proposed to improve the reproducibility of data intensive experiments. All of these tools are used at different stages of the data curation lifecycle and serve different purposes. However, they all have direct access to the data that is processed. Thus, they can provide the most detailed documentation of the experiment. We can reuse this information and feed it (semi-) automatically into DMPs. To achieve this, we need a machine-actionable data model that allows organising this information in a structured way. Currently, there is no such model that could accommodate these needs.

For this reason, we propose machine-actionable data management plans (maDMPs) that cover the same themes as standard DMPs, but particular sections are completed with information obtained from existing tools. maDMPs can be considered as automatically collected metadata about experiments. They accompany experiments from their very beginning and are updated during their course. Consecutive tools used during processing read and write data from maDMPs. maDMPs can become a universal format for exchange of (meta-) data between the systems involved in processing. For example, a workflow engine can add provenance information to the maDMP, a file format characterisation tool can supplement it with identified file formats, and a repository system can automatically pick a suitable content type for submission and later automatically identify applicable preservation strategies. maDMPs improve reproducibility of scientific experiments, because they structure and facilitate

[1] DMP Online: http://dmponline.dcc.ac.uk

[2] DMP Tool: http://dmptool.org

[3] CERN workshop on Active DMPs: http://indico.cern.ch/event/520120/attachments/1302179/2036378/CERN-ADMP-iPRES206.pdf

maintenance of automatically captured information. Researchers benefit from having less bureaucratic procedures to follow. Funders and repositories can automatically validate DMPs. For example, they can check whether the specified ORCID[4] or e-mail are correct, whether the data is available at the specified repository, and whether the data checksums are correct – in other words, whether the information reflects reality.

In this paper we present our ongoing work on realising maDMPs. We analyse existing tools from the domains of digital preservation, reproducible research, open science, and data repositories that cover the full data lifecycle. We present their mapping to the Digital Curation Centre (DCC) checklist, discuss lessons learned, and identify limits of automation and machine-actionability. Based on that, we define requirements for machine-actionable data management plans and propose a data model. The model uses ontologies and is based on current data management plan themes. It also reuses and integrates with existing domain specific standards.

Mapping

The starting point for our investigation was a hypothesis that it is possible to reuse existing information to feed it into maDMPs. This information is captured by tools included in scientific investigations and developed to support researchers during the whole data curation lifecycle, ranging from metadata dictionaries to virtualisation and containers. To organise this information in a systematic way, we need a machine-actionable data model. We presented this initial hypothesis at the RDA plenary meeting in Denver and received positive feedback[5]. In this section, we present the mapping of identified tools and standards to DMP sections (aka themes).

In our analysis we used the Digital Curation Centre checklist v4 (DCC, 2013) that is a generic template for DMPs. It is based on common funder requirements and is reused and customised by institutions that either translate or select subsets of categories. Thus, we ensure that our analysis of categories included in the majority of DMPs is comprehensive.

We also looked for tools and standards in the domains of digital preservation, reproducible research, open science, and data repositories that cover the full lifecycle of data in the following categories:

- Collaboration platforms to enable virtual collaboration, but also backups and versioning, e.g. Open Science Framework[6], Jupyter Notebook[7];
- Workflow engines to describe the data transformation process, e.g. Taverna[8], Pegasus[9];
- Provenance to provide evidence on how the experiment was conducted, e.g. PROV-O[10], OPM[11];

[4] ORCID: http://orcid.org

[5] 8th RDA Plenary meeting slides: https://www.rd-alliance.org/group/active-data-management-plans-ig/post/slides-actionable-dmps-presented-joint-meeting-denver

[6] Open Science Framework: http://osf.io

[7] Jupyter Notebook: http://jupyter.org

[8] Taverna: http://taverna.org.uk

[9] Pegasus: http://pegasus.isi.edu

[10] PROVO-O: http://w3.org/TR/prov-o/

[11] OPM: http://openprovenance.org

- Tools to track execution of experiments and to port them to other environments, e.g. CDE, PMF[12], ReproZip[13];
- Virtualisation and containers to encapsulate an experiment's environment, e.g. Vagrant[14], Docker[15], Research Objects[16], HDF5[17];
- Metadata and ontologies to provide necessary context, e.g. PREMIS[18], Dublin Core[19];
- Data repositories to store and share data, e.g. CKAN[20], Zenodo[21], Phaidra[22];
- Unique identifiers and data citation to precisely locate the data, e.g. DOI[23], recommendations of the RDA working group on Data Citation[24].

Table 1 presents the mapping of DMP sections to existing tools and standards. We are still working on its extension. The columns *Section* and *Guidance* correspond to the DCC DMP Checklist v4. Based on the guidance for each question, we identified *Keywords*. In the last column, we present the identified tools and models. Our intention was not to create a complete and finite list of tools, but to identify such that cover questions addressed by DMPs. We were able to define mappings for each section. This shows that existing standards, tools and models provide suitable information, enabling maDMPs implementation. Naturally, the maturity and uptake of each tool and model vary across disciplines. Hence, before maDMPs can be introduced, it will be necessary to modify, extend and integrate them. This may require further discussion within specific communities, but aligns with the recommendations from "Realising the European Open Science Cloud" that states:

> 'The complexity of the current data-sharing practices and mechanisms requires gentle, rather than restrictive, regulation of existing ontologies, especially across domains, with identifier mappings as practiced already in various communities" (European Commission, 2016).

[12] Process Migration Framework: http://ifs.tuwien.ac.at/dp/process/projects/pmf.html
[13] ReprопZip: http://vida-nyu.github.io/reprozip/
[14] Vagrant: http://vagrantup.com
[15] Docker: http://docker.com
[16] Research Objects: http://researchobject.org
[17] HDF5: http://support.hdfgroup.org/HDF5/
[18] PREMIS: http://loc.gov/standards/premis/
[19] Dublin Core: http://dublincore.org
[20] CKAN: http://ckan.org
[21] Zenodo: http://zenodo.org
[22] Phaidra: http://phaidra.org
[23] DOI: http://doi.org
[24] RDA Data Citation Recommendations: http://rd-alliance.org/group/data-citation-wg/outcomes/data-citation-recommendation.html

Table 1. Mapping of tools and models to sections of DCC DMP Checklist v4.

Section	Question	Keywords	Tools and Models

Administrative Data	ID, Funder, Grant Reference Number, PI / Researcher ID, Contact, Date of Last Update	administrative	**Tools:** directory services (LDAP, Active Directory) and ORCID for user information; DOI for DMPs;
			Models: FOAF, Dublin Core;

Data Collection	What data will you collect or create?	type, format, size	**Tools:** content profiling (FITS, DROID, C3PO, exiftool),risk registries (PRONOM);
			Models: PREMIS;

	How will the data be collected or created?	provenance, process description, versioning, naming convention	**Tools:** execution monitoring (CDE, PMF, reproZip), workflow engines (Taverna, Pegasus), virtualisation and containers (Docker, VBox), code repositories (GitHub); collaboration platforms (OSF), virtual environments (Jupyter);
			Models: PROV-O, OPM, Dublin Core, Context Model;

Documentation and Metadata	What documentation and metadata will accompany the data?	metadata, documentation	**Tools:** wikis (redmine, confluence, OSF, GitHub), readmes, generated documentation (nanopublications, javadoc), Docker file for Docker images;
			Models: domain specific standards (biosharing.org), Dublin Core;

Ethics and Legal Compliance	How will you manage any ethical issues?	ethics, access control	DMPs are awareness tool, text description needed (DMP Roadmap); **Models:** see access control and security;
	How will you manage copyright and Intellectual Property Rights (IPR) issues?	licenses, policies	**Tools:** EUDAT tool, PERICLES Policy editor; **Models:** Creative Commons Ontology[25]
Storage and backup	How will the data be stored and backed up during the research?	storage, backup	**Tools:** Institutional storage and cloud services (ownCloud, data centres) **Models:** new developments needed

[25] Creative Commons Ontology: https://www.w3.org/Submission/ccREL/

Section	Question	Keywords	Tools and Models
	How will you manage access and security?	access control, security	**Tools:** collaboration platforms, directory services; **Models:** Basic Access Control ontology[26]
Selection and Preservation	Which data should be retained, shared, and/or preserved?	preservation planning	**Tools and Models:** preservation planning (Plato and SCAPE project)
	What is the long-term preservation plan for the dataset?	repository, costs	**Tools:** repositories (re3data, Zenodo, Phaidra); costing tools (CCEx); **Models:** PREMIS
Data sharing	How will you share the data?	sharing, marketing, identifiers	**Tools:** repositories, data sharing platforms (GitHub, datahub, Zenodo), social media (twitter, Facebook), community portals (Researchgate, LinkedIn); **Models:** new developments needed;
	Are any restrictions on data sharing required?	embargo, legal regulations	**Tools:** DMPs are awareness tool, text description needed (DMP Roadmap); **Models:** PREMIS, Publishing Status Ontology[27];

Responsibilities and Resources	Who will be responsible for data management?	roles	**Tools:** directory services;
			Models: FOAF, Dublin Core;

	What resources will you require to deliver your plan?	resources	DMPs are awareness tool, text description needed (DMP Roadmap), applies mostly to initial DMPs.

Analysis and Requirements

In this section, we present the analysis of DMP requirements and discuss lessons learned from the mapping. We also describe limitations and derive requirements for a data model that will underpin machine-actionable data management plans.

DMP Phases

DMPs have different states and are evolving documents. When a DMP is required before the actual research starts, then it is less detailed and should be considered in

[26] Basic Access Control Ontology: http://www.w3.org/ns/auth/acl

[27] Publishing Status Ontology: http://purl.org/spar/pso

terms of a DMP proposal. At this stage, it is rather an exercise for researchers to identify any constraints on using the existing datasets and to select services enabling collaboration and backup of data during the project. In this phase, most of the information is textual; hence there is little that can be automated in terms of information sourcing and its validation. However, this textual information can be organised into a document following a precise schema, so that the respective paragraphs can be easily filtered, for example, for reporting or presentation on a web page.

When DMPs are created during the project rather than at its end, then more information is available, because the actual research and experiments are already being performed. This is when the tools discussed in the mapping are used. Hence, this is the stage in which the machine-actionability of DMPs will bring most of benefits, but there are still questions that require textual input. These are mostly questions asking for explanations about decisions taken during the project, such as reasons for using a particular dataset or standard.

DMP Relation to Data

One of the major problems of standard DMPs is their generality. They specify what kind of data is produced and where it is stored, but there are cases when there is insufficient information to enable the accessing of this data, for example, lack of DOI or access rights. A possible solution to circumvent this problem is to pack the DMPs together with the actual data, similar to the way that Research Objects do (Bechhofer et al., 2013). Research Objects can be seen as folders in which research data is organized. They also contain an ontology file that annotates each of the files within the folder structure to provide a type of metadata about each file. However, such a solution is limited, because it only applies to experiments that produce low amounts of data. It does not scale up for the big data. Hence, maDMPs should not be packaged together with data, but rather contain a precise list of data objects, their classification and precise information, allowing its unambiguous location and validation. We should consider maDMPs as metadata about scientific experiments. The link to the actual data can be established using combination of parameters such as: name, type, hash, DOI. Such information is lightweight and can be automatically captured, and validated.

DMP Openness

Despite standardisation efforts, there are still many co-existing and overlapping schemas in Bioinformatics. A similar situation can be observed in other disciplines. There is no single solution that addresses all needs at once. We cannot decide which metadata standards are best for each area. This stems from the common practice within a given research area. maDMPs must be flexible to accommodate these different needs and must reuse existing standards. For example, the PROV-O ontology is one of the provenance standards. There is no need to develop a new standard specific to DMPs, as well as to force a single one.

It is also not our role to decide how to implement an experiment, that is, we cannot say whether someone should use a workflow engine, implement a python script, and so on. We also cannot require the use of containers, such as Docker, or virtual machines. These are researchers' decisions and stem from community good practice and other recommendations, such as FAIR. Hence, the maDMPs must allow for the description of experiments with differing qualities of data management, i.e., both well and badly managed. However, the machine-actionability will ease their evaluation.

DMP Closed Questions

The manual input and textual descriptions are inevitable, because DMPs also require explanations for actions taken, for example, "Explain why you have chosen certain formats" or "If you need to restrict access to certain communities or apply data sharing agreements, explain why". However, it is possible to transform these questions into closed questions in which the user must select answers from a pre-defined list. For example, the EU commission identified three most common reasons for opt-out and not sharing data: (1) privacy, (2) intellectual property rights, (3) sharing might jeopardise project's main objective[28]. These could be used as possible answers in a questionnaire.

Similarly, other categories in which manual input is required can be converted into closed questions using questionnaires. For example, when choosing a repository in which the data is to be preserved, the user can be presented with an overview of data objects identified in previous steps and asked to select whether the given object is preserved, where, and for how long. Furthermore, based on previously provided information, such as, file format, volume, licences, the tool supporting information acquisition (e.g. DMP Roadmap) could suggest a repository that best fits the requirements. Such automation is possible when the answers are selected from a pre-defined list and not as free text.

DMP Machine-Actionability Limitations

Machine-actionability allows the automation of information collection, integration, and validation. In a perfect world, the maDMPs would enable these three actions for each of its categories, thereby almost completely reducing effort from researchers and evaluators. However, we have identified that this is not possible for each category. Below we discuss each of these actions and provide examples of what is possible, and what is not.

Automatic information collection

DMPs contain two kinds of information. First, that which describes the characteristics of actual data objects used and produced in the experiment. This is covered in categories such as *Data Collection* or *Metadata*. Second, that which describes actions and conditions that apply to this data, such as backups, versioning, or licenses.

In the first case, we can automatically collect information on the data from an environment by deploying monitoring systems, reading out embedded metadata, importing provenance traces that are captured by workflow engines, or using content profiling tools to identify file formats and size. This is possible because we can depend on tools that are either used by researchers during experiments, or by other stakeholders, such as repositories. In the second case, automatic information collection is limited. For example, it is not transparent to research tools whether data is backed up or not. This is both the case when backups are performed manually by researchers copying it to a different location, or automatically when running computations in a controlled environment. In such cases maDMPs still need to depend on manually provided information, whether necessary measures are in place. Similar considerations apply for legal compliance and ethics. Such problems can be addressed by interactive

[28] EC Open Data: https://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf

questionnaires and their output can be fed into maDMPs. An example could be the EUDAT license selector[29] that helps in when selecting a proper license.

Automatic information integration and reasoning

Machine-actionability enables querying of information for reporting or further processing. It also helps in the auto-completion of information by re-using information provided once in different contexts, as well as to infer inexplicit information. This is possible using pre-defined rules. For example, when one of the data objects described in a DMP is a GitHub repository then the system can infer that this data is versioned. Similarly, if data is located in a cloud service such as ownCloud[30] run by Geant, then a system can infer that data is backed up. This is possible only if a given system is known to a tool that processes maDMPs.

Furthermore, information on each data object can be used to automatically create aggregations and statistics such as the size of a collection, file types, licenses used, restrictions on reuse, etc. This is especially useful when DMPs are created for intermediate reporting and are snapshots of ongoing research.

The information provided in a structured way, following a vocabulary of terms, can also be useful when the data is read from a maDMP, for example, when the data is handed over to a repository. If there are any restrictions on access to data, they can be compared with the policies of a repository. Projects like SCAPE (Becker et al., 2014) and PERICLES (Biermann et al., 2016) worked on automation of policies.

Such examples show that machine-actionability does not mean that data is only automatically collected. There are cases when manual input is required, but once the information is organised following a vocabulary of terms and using a schema, then it is possible to benefit from integrations and reasoning.

Automatic information validation

Information validation can be understood as: (1) validating whether all required information was provided, (2) validating whether the provided information is not contradictory, (3) validating whether the information is true.

A precisely defined schema enables the validation of DMP completeness. This is independent of whether the information was provided manually or automatically.

The rules and reasoning described above can also be used for the validation of information. For example, we can identify that a dataset which has an embargo period and contains personal sensitive data cannot be preserved in an open access repository.

However, the main challenge lies in validating whether the provided information is true and reflects reality. As with automatic information collection, we can more easily validate much of the information describing the data itself than actions and conditions applying to the data. Hence, it is possible to check, for example, whether a user with a given ORCID exists (in a registry, not as a human), if a provided DOI links to an existing dataset, or if hashes of files match to their provenance traces. However, in categories depending on manually provided information and those in which access to the infrastructure is limited, we still have to depend on the honesty of people completing a questionnaire. For example, we are not able to determine whether a given dataset can be used in an experiment or whether the backups are really performed. In fact, this is a matter of the machine-actionability of Service Level Agreements, policies and in general trust in a supporting infrastructure.

[29] EUDAT license selector: https://ufal.github.io/public-license-selector/

[30] ownCloud: http://www.geant.org/Services/Storage_and_clouds/Pages/ownCloud.aspx

It is important to understand these limitations of maDMPs and we need to keep in mind that DMPs are not only a technical documentation of an experiment, but also an awareness tool, and this will not change when we move to maDMPs.

Requirements

Based on our observations, we formulated requirements for maDMPs presented below:

1. maDMPs must follow a precisely defined schema to enable machine actionability;
2. maDMPs must be open to be able to incorporate new data types, models and descriptions;
3. maDMPs cannot impose limits on technologies and must allow experiments to be implemented using any technology of choice;
4. maDMPs cannot be an evaluation means per se – they must be able to describe both good and bad data management practices and enable their later evaluation;
5. maDMPs must accommodate needs for manually completed information that cannot be evaded;
6. maDMPs should use closed questions whenever possible and depend on controlled vocabularies, thus reducing the need for textual descriptions;
7. maDMPs must be customisable – the principle of one-size-fits-all does not apply here. maDMPs must adapt to best practices of each domain;
8. maDMPs must be scalable – they should be able to describe both small local experiments, as well as distributed experiments. Thus, they should contain descriptions and links to data objects, but not the data itself;
9. maDMPs must link to unique and identifiable entities, such as people, repositories, and licenses thus enabling validation.

Data Model for maDMPs

In this section we propose a data model that can be used to implement maDMPs. We derived it based on the requirements defined in previous sections and analysis of information modelling techniques.

We devised a common model for maDMPs that fulfils these requirements. It is an OWL ontology which can be found online[31]. The study of information modelling techniques suggests that the hybrid approach for the integration of models is suitable for maDMPs (Wache et al., 2001). It requires a top-level vocabulary to which the models are mapped. The proposed common vocabulary is based on DMP themes[32]. Domain specific standards can integrate with it. Thus each community or research funder can

[31] Common model for maDMPs: http://purl.org/madmps

[32] Revised DMP themes: http://www.dcc.ac.uk/sites/default/files/documents/tools/dmpOnline/DMP-themes-FINAL-Dec2016.pdf

devise their own maDMP implementations that best suit their needs and practices. Similar to standard DMPs, the implementation requires identifying the required scope of information and the existing standards that best reflect community needs.

Figure 1 presents an example of an instance of the model. It contains a root element *Data Management Plan* that links to two data objects that are described by a maDMP: *File* and *Source Code*. *Data Object* is a generic class that can be further specialised using the right term from a controlled vocabulary of object types, for example, *File*, *Source Code*, *Compiled Software*, *Container*, *Virtual Machine*, and so on. For each of the data objects one must define object properties that reflect DMP themes. For example, *has Metadata* object property links to *Metadata* about the *File*. The specific metadata information is modelled using domain specific standard and is not depicted in Figure 1. If an object property is not defined for a given object, then this is a clear indicator that a DMP is incomplete. Each of the objects representing themes contains a data property with information on a standard that is used as an extension. This is information for machines enabling automatic selection of the right parser for reading out information.

The proposed architecture that uses a common model and requires domain specific extensions will not result in a multitude of incompatible maDMP solutions, because all of them have a common vocabulary that can be seen as a common interface for accessing information about an experiment. Such a common interface enforces that all sections of DMPs are covered, or are intentionally omitted if not applicable in a given domain. Furthermore, the common model is extensible and can adapt to changes. For example, containers such as Docker are becoming one of the means of increasing reproducibility of experiments. They can also be considered as data objects that have to be covered by DMPs. An extension to the controlled vocabulary of DMP terms would enable this without having impact on other DMPs.

In Miksa et al. (2016) we have shown on a use case from biomedical domain how information on file formats, provenance, technical dependencies, and validation requirements can be automatically collected and integrated into a common model using extension ontologies. Figure 2 depicts an excerpt of such a model. The maDMPs can be modelled in a similar way.

The proposed model is a technical solution needed to realize maDMPs and the researchers do not have to interact with it any way. It is an important building block, but other developments are needed, such as pilot studies in different domains to identify domain specific extensions to the core model and necessary tools integrations.

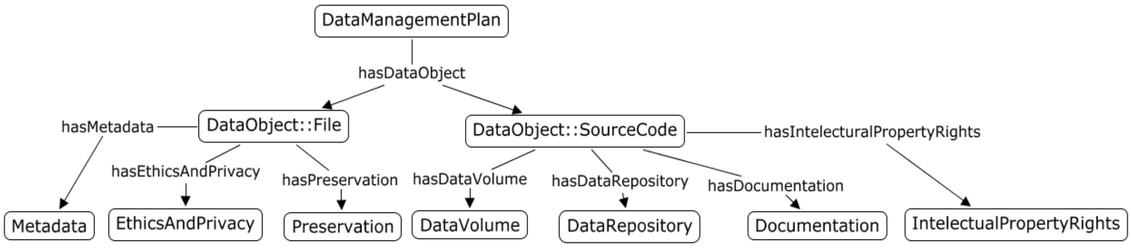

Figure 1. Example of the proposed common model for machine actionable DMPs.

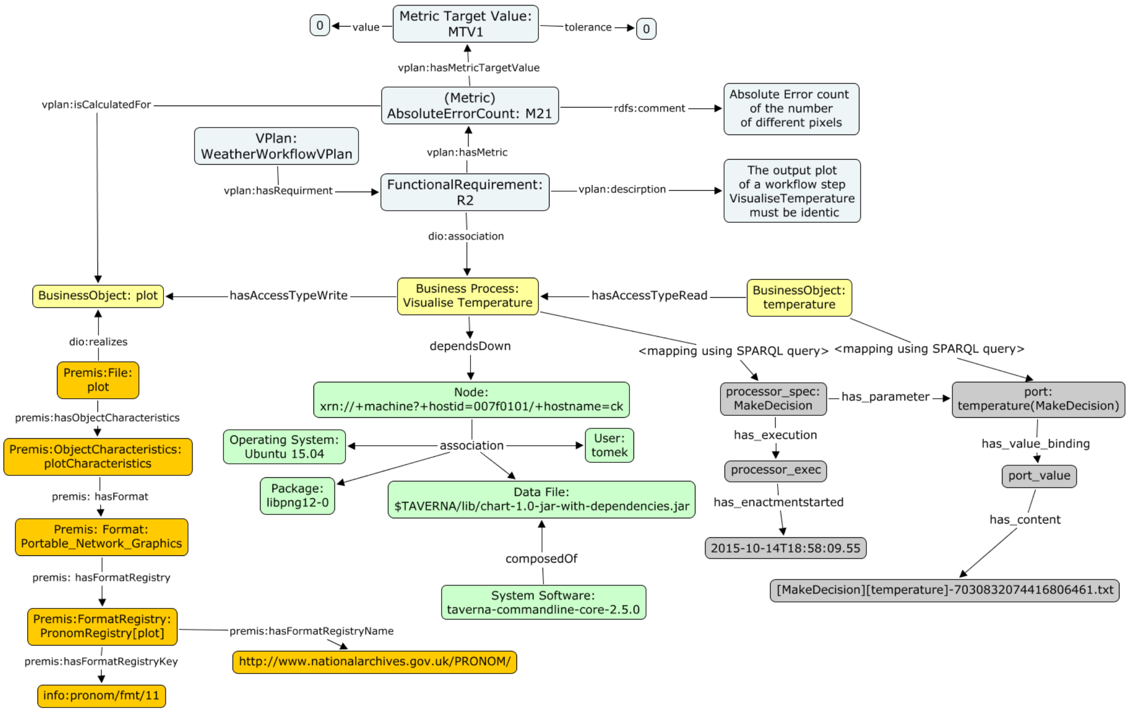
0
value
Metric Target Value: MTV1
tolerance
0
vplan:hasMetricTargetValue
vplan:isCalculatedFor
(Metric) AbsoluteErrorCount: M21
rdfs:comment
Absolute Error count of the number of different pixels
VPlan: WeatherWorkflowVPlan
vplan:hasMetric
vplan:hasRequirment
FunctionalRequirement: R2
vplan:descirption
The output plot of a workflow step VisualiseTemperature must be identic
dio:association
BusinessObject: plot
hasAccessTypeWrite
Business Process: Visualise Temperature
hasAccessTypeRead
BusinessObject: temperature
dio:realizes
dependsDown
<mapping using SPARQL query>
<mapping using SPARQL query>
Premis:File: plot
Node: xrn://+machine?+hostid=007f0101/+hostname=ck
processor_spec: MakeDecision
has_parameter
port: temperature(MakeDecision)
premis:hasObjectCharacteristics
Operating System: Ubuntu 15.04
association
User: tomek
has_execution
has_value_binding
Premis:ObjectCharacteristics: plotCharacteristics
processor_exec
port_value
premis: hasFormat
Package: libpng12-0
Data File: $TAVERNA/lib/chart-1.0-jar-with-dependencies.jar
has_enactmentstarted
has_content
Premis: Format: Portable_Network_Graphics
2015-10-14T18:58:09.55
composedOf
premis: hasFormatRegistry
[MakeDecision][temperature]-7030832074416806461.txt
System Software: taverna-commandline-core-2.5.0
Premis:FormatRegistry: PronomRegistry[plot]
premis:hasFormatRegistryName
premis:hasFormatRegistryKey
http://www.nationalarchives.gov.uk/PRONOM/
info:pronom/fmt/11

Figure 2. Example of integrated ontologies for describing file formats (orange), technical dependencies (green), provenance (dark grey), and validation requirements (light grey) of a scientific workflow (yellow) (Miksa et al., 2016).

Benefits and New Opportunities

In this section we discuss what benefits maDMPs bring to various stakeholders and what new opportunities arise when maDMPs are introduced.

FAIR maDMPs

We can consider maDMPs as metadata about scientific experiments. maDMPs follow a formally defined schema and provide a rich description of data objects used and created during research. They also contain references to existing datasets. maDMPs enable automatic generation of landing pages that summarise experiment descriptions and provide a single point of information on an experiment. Each maDMP can have its own unique identifier. Thus, they support each of the FAIR principles.

maDMPs Repository

Currently, researchers are encouraged to publish their DMPs in selected journals. When searching for relevant DMPs they use text based search using keywords. The machine-actionability of maDMPs gives more search options by using similarity metrics that identify DMPs which use similar resources, standards, technologies, or require similar infrastructure, skills, or budget for implementation.

Evaluation of Data Management Practices

For reviewers and research evaluators, the criteria on quality for DMPs and experiment reproducibility are highly subjective. If it happens that data is provided in an open repository like Zenodo, but does not have proper organisation and metadata description that would enable its reuse, the data may meet the requirements of publishers because it has a DOI, which is the only thing that can be easily checked manually by reviewers. This is because there are no quantitative measures that could assess how good a DMP is or how replicable the experiment is. This has a direct impact on the reuse of experiments, because the reusing researchers do not know how to estimate the time, computational resources and skills needed for re-execution. Current research focuses on estimating costs and resources needed for sharing and digital preservation of objects from the perspective of data producers and repository managers (4C Consortium, 2015), but there is no research on the costs and resource needed for re-execution of such reused experiments.

The automatically collected information represented in a machine-actionable data model can become a basis for further evaluation of replicability of experiments. We can devise new metrics that will enable quantification of replicability, taking into account ease of reuse, costs, computation power required, availability of resources, skills required, portability, and so on. They will cover both functional and non-functional aspects. Thus, we will be able to benchmark existing approaches and classify in what way they support replicability.

Conclusion

Data Management Plans are compulsory documents describing the data used and produced in scientific experiments. They are free-form text documents whose quality depends on both expertise and scrupulousness of people writing them. The complexity of data-driven experiments requires precise descriptions of tools and datasets used in computations to enable their reproducibility and reuse. Hence, there is a need for automatically created machine-actionable description of experiments that could improve the quality of data management plans and reduce effort of their preparation without influencing research practices. This problem was recognized by the research community and is being discussed in conferences and venues such as Research Data Alliance.

In this paper we presented our ongoing work on realising the concept of machine-actionable data management plans. We analysed existing tools from the domains of digital preservation, reproducible research, open science, and data repositories that cover the full data lifecycle. We identified their mapping to the Digital Curation Centre checklist, which is an aggregation of funders' requirements and serves as a generic template for data management plans. We presented lessons learned from the mapping and described limitations of automation and machine-actionability. As a result, we defined the requirements for machine-actionable data management plans. Furthermore, we proposed a data model that can underpin the machine-actionable data management plans, enabling flexible information integration. The model uses ontologies and is based on current data management plan themes. It can be extended with existing domain specific standards. We also provided a discussion of the benefits and new opportunities represented by maDMPs.

The future and ongoing developments focus on engaging with stakeholders from various communities to run pilot studies in which the proposed common model is extended with the standards identified together with community experts. This will require mapping of standards and software engineering tasks. Thus, we will evaluate the proposed approach. We will report back to the community through Research Data Alliance interest and working groups. We believe that the success of machine-actionable data management plans requires the joint community effort and depends on broad acceptance. For this reason we monitor the ongoing work on this and similar topics and look forward to new cooperation.

References

4C Consortium. (2015). Roadmap: Investing in curation: A shared path to sustainability. Retrieved from http://www.4cproject.eu/documents/Roadmap%20-%20V1.02%20-%2020Feb2015.pdf

Bechhofer, S., Ainsworth, J., Bhagat, J., Buchan, I., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Goble, C., Michaelides, D., Missier, P., Owen, S., Newman, D., De Roure, D., & Sufi, S. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems* 29(2), pp 599-611. doi:10.1016/j.future.2011.08.004

Becker, C., Faria, L., & Duretec, K. (2014). Scalable decision support for digital preservation OCLC systems and services. *International Digital Library Perspectives, 30*(4), pp 249 – 28. doi:10.1108/OCLC-06-2014-0025

Biermann, J., Eggers, A., Corubolo, F., & Waddington, S. (2016). An ontology supporting planning, analysis, and simulation of evolving digital ecosystems. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems* (MEDES). ACM, New York, NY, USA. doi:10.1145/3012071.3012081

DCC. (2013). *Checklist for a data management plan. v.4.0*. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/data-management-plans

European Commission. (2016). *Realising the European Open Science Cloud (EOSC)*. First report and recommendations of the Commission High Level Expert Group. doi:10.2777/940154

Miksa, T., Rauber, A., & Mina, E. (2016). Identifying impact of software dependencies on replicability of biomedical workflows. *Journal of Biomedical Informatics 64C,* pp. 232-254. doi:10.1016/j.jbi.2016.10.011

Wache, H., et al. (2001). Ontology-based integration of information: A survey of existing approaches. *IJCAI'01 Workshop. on Ontologies and Information Sharing.*

Wilkinson, M., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* (160018). doi:10.1038/sdata.2016.18