Research Object Crates and Machine-actionable Data Management Plans

Tomasz Miksa[1 [0000−0002−4929−7875]], Maroua Jaoua[1 [0000−0001−8109−9644]], and Ghaith Arfaoui[1 [0000−0001−9554−5684]]

TU Wien & SBA Research, Austria

Abstract. Research Object Crates package research data with its metadata. Machine-actionable Data Management Plans describe what happens to data over the research lifecycle. To automate data management tasks and to lessen the burden on researchers, we investigated how information can be exchanged between them. In this paper we present a mapping and evaluate it on a set of existing machine-actionable Data Management Plans and Research Object Crates. Results show that a significant part of information can be exchanged and thus automation of data management tasks is possible.

Keywords: maDMPs · RO-Crate · FAIR · data management · automation

1 Introduction

Research Data Management (RDM) should not be a sole responsibility of researchers, because of its complexity and time overhead. Researchers not only need guidance on how to make their research outputs FAIR [11], but also tools that relive them from time-consuming data management tasks. For example, they need tools that help in capturing metadata about produced data, or tools that pre-fill Data Management Plans (DMPs) based on existing information.

There is a number of solutions proposed that try to address challenges of describing scientific data and its management using structured information. Many of them use semantic technologies and depend on Linked Open Data (LOD). We believe that by exchanging and reusing information captured at different stages of research data lifecycle, we can automate data management tasks and thus lessen the burden on researchers. In this paper we investigate Research Object Crates (RO-Crates) and Machine-actionable DMPs (maDMPs) to show how this can be achieved.

RO-Crate [9] is a lightweight approach to packaging research data with its metadata. It is based on schema.org annotations in JSON-LD, and aims to make metadata description accessible and practical for use in a variety of situations, from an individual researcher working with a folder of data, to large data-intensive computational research environments.

MaDMPs are an emerging standard for exchange of DMPs between systems involved in research data management. They were developed within the DMP

Common Standards Working Group of the Research Data Alliance and are its official output [8]. A substantial part of the maDMPs follows the W3C DCAT [2] specification.

In this paper we present how information can be exchanged between RO-Crates and maDMPs to automate and ease management of research data. To do that, we perform a mapping between RO-Crates and maDMPs. Furthermore, we present a tool that generates: (1) RO-Crates using information from maDMPs, (2) fill in maDMPs using RO-Crates . We evaluated our approach by migrating a selection of publicly available RO-Crates and maDMPs.

The paper is structured as follows. Section 2 presents related work. Section 3 describes the use cases and explains the rationale for exchange of information between RO-Crates and maDMPs. Section 4 presents the mapping. Section 5 presents the migration tools and discuss the results of migration. The last section presents conclusion and future work.

2 Related work

Data management plans are a common requirement of funding bodies and institutions. DMPs describe the data that is used and produced during research, where the data will be stored, which licenses apply, and to whom credit should be given [7]. Researchers can use tools to create DMPs. These tools provide questionnaires that must be answered to create a DMP complaint with a selected funder template. A list of most recent DMP tools can be found in [4].

The Research Data Alliance (RDA) DMP Common Standards[1] working group developed an application profile for maDMPs [8] that provides machine-actionable representation of information contained in DMPs consisting of atomised, structural data. Breaking the information down into specific fields creates added value to all stakeholders in the research data lifecycle such as researchers and funders, but also data stewards, repository operators, etc. who can provide and reuse information using systems acting automatically on their behalf [7]. The application profile was developed in an open and consensus-driven manner [10], [6], [5].

MaDMPs reuse concepts such as *Dataset* or *Distribution* from the W3C DCAT specification. The application profile can be serialized to JSON, but there is an ongoing work on a semantic web representation of maDMPs. The full specification of the application profile can be found online[2].

RO-Crate builds on top of Research Objects (ROs). The aim of ROs was to replace traditional academic publications of static PDFs, with a complete and structured inventory of items that contributed to the research outcome, including their identifiers, provenance, relations and annotations [3]. ROs combine existing Linked Data standards: W3C RDF, JSON-LD, OAI-ORE, W3C Web Annotations, PROV, Dublin Core Terms, ORCID.

[1] https://www.rd-alliance.org/groups/dmp-common-standards-wg
[2] https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard

RO-Crate packages research data with their structured metadata, based on schema.org annotations in a formalized JSON-LD format that can be used independent of infrastructure to encourage FAIR [11] sharing of reproducible datasets and analytical methods. RO-Crate is also an initiative that aims at bringing together data repository operators with existing Research Object, workflow and provenance communities [9]. Details on the RO-Crate can be found in the open repository[3].

3 Use cases

In this section we describe two scenarios for exchange of information between RO-Crates and maDMPs. We explain the rationale, goals and benefits. Both use cases are illustrated in Figure 1.

3.1 maDMPs to RO-Crates

MaDMPs are living documents updated throughout the research lifecycle. There are settings, e.g. National Science Foundation (NSF) funded grants, when the maDMPs are written before project starts. In such cases, maDMPs describe what data will be produced, where it will be kept, etc. In other words, maDMPs describe future actions, e.g. such maDMPs can state that source code of simulation will be written in Python and will be shared on GitHub using MIT license.

RO-Crates do not exist at this early stage, but information included in maDMPs can be used to facilitate its creation. For example, datasets specified in the maDMP can be used to bootstrap creation of RO-Crates , by providing initially pre-filled RO-Crates that can later be edited by researchers, either manually or using other tools that add information on processing. In this use case (upper part of Figure 1), we aim to identify which maDMP concepts and fields can be used to generate RO-Crates .

3.2 RO-Crates to maDMPs

MaDMPs should be updated when researchers come to a specific stage of research, e.g. when they write a publication and share their data, or when the research project ends. In such cases, the maDMPs describe the existing data and actions that were already performed, e.g. simulation data was uploaded to an open repository and is available under CC-BY license.

When the data already exists, then RO-Crates describing the data can also be created. RO-Crates can be either manually created using editors, or automatically exported from tools that support RO-Crate integration. For example, *Workflow Hub*[4] supports exporting *Galaxy* [1] workflows as RO-Crates . RO-Crates can later be used to fill in parts of maDMPs and thus reduce workload imposed on researchers.

[3] https://github.com/ResearchObject/ro-crate
[4] https://workflowhub.eu/

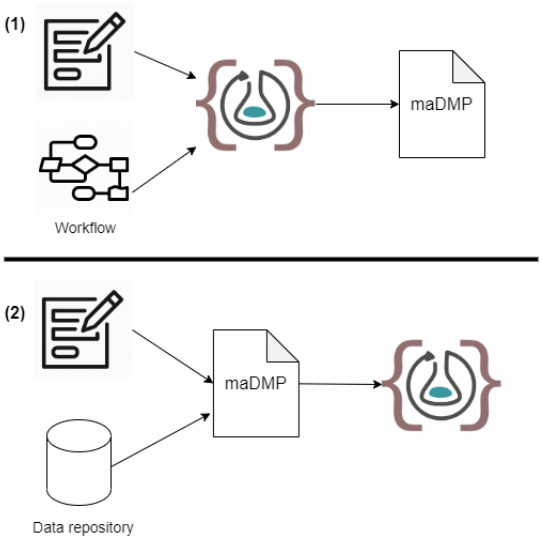

Fig. 1. Use cases illustrated: (1) information used in RO-Crates is re-purposed for maDMPs, (2) maDMPs are used to generate RO-Crates.

In this use case (lower part of Figure 1), our goal is to identify which RO-Crate concepts and fields can be used to fill in maDMPs.

4 Mapping

In this section we describe how we did the mapping between maDMPs and RO-Crates . We provide en example of mapped properties, statistics on the mapping completeness and describe assumptions made.

4.1 Methodology

We performed the mapping in two iterations. In the first iteration, we investigated RO-Crates and looked for corresponding concepts in maDMPs. Each RO-Crate includes information about one dataset. However, a maDMP has a broader focus and may have multiple datasets. Therefore, multiple RO-Crates can be translated into one maDMP, if they are part of the same project.

In the second iteration, we took maDMPs as the staring point and looked for matching concepts in RO-Crates . Similarly to the first iteration, we checked all the properties of maDMPs to find matching ones in RO-Crate. The specification of RO-Crates mentions the possibility to use schema.org metadata to supplement RO-Crates and by other Linked Data Vocabularies when properties are missing. Therefore, attributes which are present in maDMPs and missing in RO-Crates are accounted for by other Linked Data Vocabularies. Since maDMPs may include multiple datasets, one maDMP can generate multiple RO-Crates .

The full mapping can be found in the GitHub repository[5]. The mapping shows: mapped properties, unmapped properties of maDMPs, and unmapped properties of RO-Crates . Table 1 is an excerpt from the mapping table.

Table 1. Excerpt from the mapping table found in the the tool's GitHub repository.

RO-Crate property	@type of RO-Crate property	maDMP property	Parent of maDMP property	Assumption
cost	Dataset	cost	dmp	In DMP, the cost represents a list of costs related to data management. However, the cost for RO-Crate may not include all costs.
costCurrency	cost	currency_code	cost	This is not explicitly mentioned in RO-Crate website. But, cost properties can be found in jsonld context used for RO-Crates .
description	Dataset	description	dataset	
email	ContactPoint	mbox	contact	
license	Dataset/File	license	distribution	
@id	Grant	identifier	funder_id	

4.2 Statistics

Figure 2 gives an overview about the statistics of mapped properties. We managed to map in total 46 properties. However, 7 of the mapped attributes are not exact, that is, certain assumptions had to be made (see below). The remaining 39 are exactly mapped. There are 33 properties of maDMPs and 13 properties of RO-Crates that we did not manage to map. 9 of the unmapped maDMP properties do not need to be mapped at all since the children of these properties are mapped. 4 of the unmapped RO-Crate properties do not need to be mapped since they are specific to the format of JSON-LD files, for example, @context property is specific to files which have JSON-LD format. Since RO-Creates are based on schema.org, not all possible properties are mapped, because there are too many of them. We focused only on those schema.org properties that are used in the documentation of RO-Crates.

4.3 Assumptions

There is a significant overlap between the RO-Crates and madmps. However, the mapped properties are not always exact since both concepts have sometimes different definitions. Therefore, we had to make a number of assumptions.

[5] https://github.com/GhaithArf/ro-crate-rda-madmp-mapper/blob/master/README.md

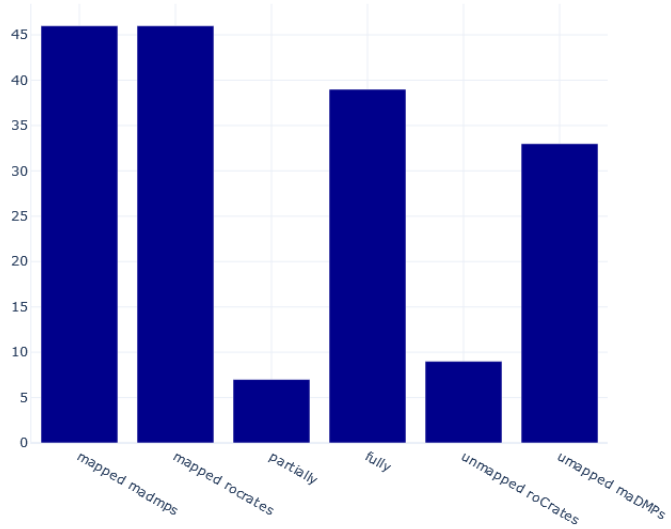

Fig. 2. Overview about the mapped and unmapped maDMP and RO-Crate properties.

(a) RO-Crate includes one dataset. The latter can have other nested datasets. We assume that the properties for the dataset at the root of a RO-Crate are equivalent to the properties of maDMP. For example, we assume that the contact person for the DMP is the same as the contact person of the RO-Crate's dataset.
(b) Nested datasets in the RO-Crates are included as elements of the property distribution in maDMPs.
(c) RO-Crates can include deeply nested datasets. These datasets are not taken into consideration while mapping. Only the root dataset and one sub-dataset can be mapped.
(d) RO-Crates have a flat JSON-LD file format. They require the property @id for each entity. However, an identifier is not always present for entities of maDMP. In case of absence of identifiers, we assume that the title of maDMP properties represents @id property in RO-Crate entities.
(e) In maDMP, the cost represents a list of costs related to data management. However, the cost in RO-Crate may not include all costs. We assume the cost property to be partially equivalent.
(f) Two properties from both standards have the same definition and different formats. For instance, the property language from maDMP is expressed using ISO 639-3. However, the language property of RO-Crate does not necessarily follow the same convention. These properties are still assumed to be equivalent. However, it is required to adjust the format manually.

(g) There are properties that are automatically generated. An example is the modification date of the maDMP. This property is not present in RO-Crates. But, it corresponds to the date of creation of the maDMP and can be automatically filled.

There are also properties that are not mapped for the following reasons.

(a) Some properties are not covered in both concepts. For instance, almost all properties which have to do with quality assurance, privacy, ethics and security are missing in RO-Crate and cannot be translated.
(b) The property dmp is an important property for maDMP. But, it cannot be mapped since RO-Crate is an approach to package research data with their metadata and maDMP considers a broader concept.
(c) Some parent properties do not have an equivalent. However, their children have equivalent properties. The parent properties are not mandatory anymore since they are accounted for by their children. For example, the parent property "contact_id" has the child property "identifier". If the property "identifier" is mapped, the property "contact_id" does not need to be mapped since RO-Crates and maDMPs have different formats.

5 Evaluation and Discussion

We used publicly available examples that represent a variety of realistic scenarios and allow evaluating the mapping under real-life conditions:

- 5 maDMPs from the official RDA DMP Common Standard repository[6]
- 5 maDMPs from the Data Stewardship community on Zenodo containing examples of maDMPs[7]
- 3 RO-Crates from Research Object Crate website[8]
- 2 manually filled RO-Crates[9]

We developed a migration tool[10] to perform the conversion between both representations. The tool allows generating RO-Crate(s) from maDMP (1 to many) and maDMP from RO-Crate(s) (many to 1). The structure of both standards in addition to the previously defined one-to-one mappings between properties is defined in a JSON file. Migration is then based on that file where metadata records from one standard are extracted and structured in the way defined by the other standard.

When transforming RO-Crates to a maDMP, each metadata file is separately converted and then all of them are merged together. The structure of properties that are specific to maDMP are also defined and need to be manually filled.

[6] https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/tree/master/examples/JSON
[7] https://zenodo.org/communities/tuw-dmps-ds-2020/
[8] https://data.research.uts.edu.au/examples/ro-crate/0.2/
[9] https://github.com/GhaithArf/ro-crate-rda-madmp-mapper/tree/master/examples/rocrate
[10] https://github.com/GhaithArf/ro-crate-rda-madmp-mapper

Generating RO-Crate(s) from maDMP allows the creation of one metadata file for each dataset. For elements defined in the list of datasets whithin maDMP, information is extracted and a RO-Crate file is generated based on that. Information related to quality, privacy, ethics, and security are not migrated since no RO-Crate properties are available for them.

Similarly, some details are not migrated when converting RO-Crate(s) to maDMP. Detailed information about single files of a dataset are not fully migrated due the difference in concepts between has_part from RO-crate and distribution from maDMP. Parts of a dataset with deeply nested structure are not considered during migration and only information about the root folder representing a dataset part are taken into account. Furthermore, the nested nature of maDMP structure makes the generated metadata files more compact compared to RO-Crate, especially when datasets share the same properties.

Listing 1.1. presents an RO-Crate generated using information from the maDMP that is presented in Listing 1.2. We can observe that information on datasets and authors can be almost completely exchanged between the representations. This shows that although the full alignment between the specifications is not possible, the fields that overlap allow for meaningful conversion and can facilitate automation of data management tasks.

Listing 1.1. Example of a generated RO-Crate based on maDMP.

```
{
  {
    "@id": "https://orcid.org/0000-0001-8109-9644",
    "email": "maroua.jaoua@student.tuwien.ac.at",
    "name": "Maroua Jaoua",
    "@type": "ContactPoint"
  },
  {
    "contactPoint": {
      "@id": "https://orcid.org/0000-0001-8109-9644"
    },
    "identifier": "10.5281/zenodo.3770405",
    "description": "Data which includes the data generated by
        running the jupyter notebook",
    "hasPart": [
      {
        "@id": "https://creativecommons.org/licenses/by-nc-sa
            /3.0/igo/"
      }
    ],
    "datePublished": "2020-03-25",
    "name": "generated data",
    "Language": "eng",
    "@type": "Dataset",
    "@id": "./"
  }
}
```

Listing 1.2. Part of the original maDMP used to generate the RO-Crate.

```
{
  "contact":{
    "mbox":"maroua.jaoua@student.tuwien.ac.at",
    "name":"Maroua Jaoua",
    "contact_id":{
      "identifier":"https://orcid.org/0000-0001-8109-9644",
      "type":"orcid"
    }
  },
  "language":"eng",
  "ethical_issues_exist":"no",
  "dataset":[
    {
      "title":"generated data",
      "description":"Data which includes the data generated
          by running the jupyter notebook",
      "type":"document",
      "issued":"2020-03-25",
      "dataset_id":{
        "identifier":"10.5281/zenodo.3770405",
        "type":"doi"
      }
    ]
  }
}
```

6 Conclusion

In this paper we discussed how RO-Crates and maDMPs can be mapped to enable exchange of information between them.

The results show that a significant number of properties can be mapped directly. Thus, it is possible to pre-fill maDMPs with information on already existing data, as well as, to bootstrap creation of new RO-Crates by reusing information on planned datasets from maDMPs.

Such automated mapping can help researchers to exchange information between RO-Crates and maDMPs efficiently and without major effort.

The future work will focus on further automation of RO-Crate and maDMP creation. We plan to evaluate further examples, as well as to extend the conversion tool with a support for maDMPs serialized using ontologies.

Acknowledgment

This research was also carried out in the context of the Austrian COMET K1 program and publicly funded by the Austrian Research Promotion Agency (FFG) and the Vienna Business Agency (WAW).

References

1. Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., Blankenberg, D.: The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research **46**(W1), W537–W544 (05 2018). https://doi.org/10.1093/nar/gky379
2. Archer, P.: Data catalog vocabulary (dcat) (w3c recommendation). Online (January 2014), https://www.w3.org/TR/vocab-dcat/
3. Bechhofer, S., Ainsworth, J., Bhagat, J., Buchan, I., Couch, P., Cruickshank, D., Roure, D.D., Delderfield, M., Dunlop, I., Gamble, M., Goble, C., Michaelides, D., Missier, P., Owen, S., Newman, D., Sufi, S.: Why linked data is not enough for scientists. In: 2010 IEEE Sixth International Conference on e-Science. pp. 300–307 (2010)
4. Jones, S., Pergl, R., Hooft, R., Miksa, T., Samors, R., Ungvari, J., Davis, R.I., Lee, T.: Data management planning: How requirements and solutions are beginning to converge. Data Intelligence **2**(1-2), 208–219 (2020). https://doi.org/10.1162/dint_a_00043, https://doi.org/10.1162/dint_a_00043
5. Miksa, T., Cardoso, J., Borbinha, J.L.: Framing the scope of the common data model for machine-actionable data management plans. In: Abe, N., Liu, H., Pu, C., Hu, X., Ahmed, N.K., Qiao, M., Song, Y., Kossmann, D., Liu, B., Lee, K., Tang, J., He, J., Saltz, J.S. (eds.) IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018. pp. 2733–2742. IEEE (2018). https://doi.org/10.1109/BigData.2018.8622618, https://doi.org/10.1109/BigData.2018.8622618
6. Miksa, T., Neish, P., Walk, P., Rauber, A.: Defining requirements for machine-actionable data management plans. In: McGovern, N., Whiteside, A. (eds.) Proceedings of the 15th International Conference on Digital Preservation, iPRES 2018, Boston, MA, USA, September 24-28, 2018 (2018), https://hdl.handle.net/11353/10.923628
7. Miksa, T., Simms, S., Mietchen, D., Jones, S.: Ten principles for machine-actionable data management plans. PLOS Computational Biology **15**(3), 1–15 (03 2019). https://doi.org/10.1371/journal.pcbi.1006750, https://doi.org/10.1371/journal.pcbi.1006750
8. Miksa, T., Walk, P., Neish, P.: RDA DMP Common Standard for Machine-actionable Data Management Plans (2019). https://doi.org/10.15497/rda00039
9. O Carragain, E., Goble, C., Sefton, P., Soiland-Reyes, S.: A lightweight approach to research object data packaging. Bioinformatics Open Source Conference (BOSC2019) (Jun 2019). https://doi.org/10.5281/zenodo.3250687
10. 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
11. 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)