

# Automating Evaluation of Machine-Actionable Data Management Plans with Semantic Web Technologies

TU Wien, SBA Research

Vienna, Austria

**Raffael Foidl**

[raffael.foidl@gmail.com](mailto:raffael.foidl@gmail.com)

**Lea Salome Brugger**

[leasalome.brugger@gmail.com](mailto:leasalome.brugger@gmail.com)

**Tomasz Miksa**

[TMiksa@sba-research.org](mailto:TMiksa@sba-research.org)

# Agenda

1. Motivation
2. Methodology
3. Results
4. Discussion

# Problem

- no tool or standard procedure for assessing maDMPs
- manual assessment necessary
  - error-prone
  - time-consuming
- solution: (semi-)automate this process
  - SPARQL queries
  - helps reviewers to evaluate maDMPs
  - helps researchers to verify maDMPs

# Science Europe Evaluation Rubric

6 DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES		
Guidance for Researchers	Sufficiently Addressed The DMP...	Insufficiently Addressed The DMP...
<p><b>6a</b></p> <p><b>Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?</b></p> <ul style="list-style-type: none"> <li>Outline the roles and responsibilities for data management/stewardship activities for example data capture, metadata production, data quality, storage and backup, data archiving, and data sharing. Name responsible individual(s) where possible.</li> <li>For collaborative projects, explain the co-ordination of data management responsibilities across partners</li> <li>Indicate who is responsible for implementing the DMP, and for ensuring it is reviewed and, if necessary, revised.</li> <li>Consider regular updates of the DMP.</li> </ul>	<ul style="list-style-type: none"> <li>Clearly outlines the roles and responsibilities for data management/stewardship (for example data capture, metadata production, data quality, storage and backup, data archiving, and data sharing), naming responsible individual(s) where possible.</li> <li>Clearly indicates who is responsible for day-to-day implementation and adjustments to the DMP.</li> <li>Explains, for collaborative projects, the co-ordination of data management responsibilities across partners.</li> </ul>	<ul style="list-style-type: none"> <li>Does not discuss responsibility for data management/stewardship activities and/or does not indicate who is responsible for day-to-day implementation and adjustments to the DMP.</li> <li>Provides no description, in case of a collaborative project, on how data management responsibilities will be co-ordinated across partners.</li> </ul>
<p><b>6b</b></p> <p><b>What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?</b></p>	<ul style="list-style-type: none"> <li>Provides clear estimates of the resources and costs (for example storage costs, hardware, staff time, costs of preparing data for deposit, and repository charges) that will be dedicated to data management and ensuring that data will be FAIR and describes how these costs will be covered. Alternatively, there is a statement that no additional resources are needed.</li> </ul>	<ul style="list-style-type: none"> <li>Provides no answer or is vague about the resources required for data management and ensuring that data will be FAIR (for example resources are not listed or costed inappropriately), and/or does not describe how the costs will be covered.</li> </ul>

[1] Science Europe: Practical Guide to the International Alignment of Research Data Management – Extended Edition (Jan 2021), <https://doi.org/10.5281/zenodo.4915862>

# RDA DMP Common Standard

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "$id": "https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/tree/master/examples/JSON/JSON-schema/1.1",
  "title": "RDA DMP Common Standard Schema",
  "description": "JSON Schema for the RDA DMP Common Standard",
  "type": "object",
  "properties": {
    "dmp": {
      "$id": "#/properties/dmp",
      "type": "object",
      "title": "The DMP Schema",
      "properties": {
        "contact": {
          "$id": "#/properties/dmp/properties/contact",
          "type": "object",
          "title": "The DMP Contact Schema",
          "properties": {
            "contact_id": {
              "$id": "#/properties/dmp/properties/contact/properties/contact_id",
              "type": "object",
              "title": "The Contact ID Schema",
              "properties": {
                "identifier": {
                  "$id": "#/properties/dmp/properties/contact/properties/contact_id/properties/identifier",
                  "type": "string",
                  "title": "The DMP Contact Identifier Schema",
                  "examples": ["https://orcid.org/0000-0003-0644-4174"]
                },
                "type": {
                  "$id": "#/properties/dmp/properties/contact/properties/contact_id/properties/type",
                  "type": "string",
                  "enum": [
                    "orcid",
                    "isni",
                    "openid",
                    "other"
                  ],
                  "title": "The DMP Contact Identifier Type Schema",
                  "description": "Identifier type. Allowed values: orcid, isni, openid, other",
                  "examples": ["orcid"]
                }
              }
            }
          }
        }
      }
    }
  }
}
```

[2] Miksa, T., Walk, P., Neish, P.: RDA DMP Common Standard for Machine-actionable Data Management Plans (2020). <https://doi.org/10.15497/rda00039>

## Towards semantic representation of machine-actionable Data Management Plans

João Cardoso<sup>1</sup>[0000–0003–0057–8788], Leyla J. Garcia<sup>3</sup>[0000–0003–3986–0510], Fajar J. Ekaputra<sup>2</sup>[0000–0003–4569–2496], Marie C. Jacquemot<sup>4</sup>[0000–0002–6316–1472], Tomasz Miksa<sup>2</sup>[0000–0002–4929–7875], and José Borbinha<sup>1</sup>[0000–0001–5463–8438]

<sup>1</sup> Universidade de Lisboa, Instituto Superior Técnico & INESC-ID, Lisboa, Portugal

<sup>2</sup> TU Wien & SBA Research, Vienna, Austria

<sup>3</sup> ZB MED Information Centre for Life Sciences, Cologne, Germany

<sup>4</sup> DMP OPIDoR & Inist-CNRS, Nancy, France

**Abstract.** The concept of Data Management Plan (DMP) has emerged as a fundamental tool to help researchers through the systematical management of data. The Research Data Alliance DMP Common Standard (DCS) working group developed a core set of universal concepts characterising a DMP, in the pursuit of having making a DMP as a machine-actionable information artefact. This paper reports on the community effort to create the DMP Common Standard Ontology (DCSO), a serialisation of the DCS core set of universal concepts, giving particular focus to detailing the components of that ontology.

**Keywords:** Data Management Plan · Machine-Actionable Data Management Plan · Semantic Technology.

[3] Cardoso, J., Garcia, L., Ekaputra, F., Jacquemot, M., Miksa, T., Borbinha, J.: Towards semantic representation of machine-actionable Data Management Plans. In DaMaLOS – First Workshop on Data and Research Objects Management for Linked Open Science: Co-located at the International Semantic Web Conference ISWC 2020. PUBLISSO (Nov 2020), <https://doi.org/10.4126/FRL01-006423289>

# Mapping

- requirements from Science Europe evaluation rubric
- RDA DMP Common Standard JSON schema
- query respective fields
- ASK and SELECT

## GENERAL INFORMATION

### Administrative information

- Provide information such as name of applicant, project number, funding programme, version of DMP.

[1] Science Europe: Practical Guide to the International Alignment of Research Data Management – Extended Edition (Jan 2021), <https://doi.org/10.5281/zenodo.4915862>

```
SELECT ?title ?author ?email ?created ?language ?dmpId ?
      dmpIdType WHERE {
  ?maDMP dcso:hasContact ?contact ;
        dcso:hasDMPId ?dmp ;
        dct:created ?created ;
        dcso:language ?language ;
        dct:title ?title .
  OPTIONAL { ?maDMP dcso:hasProject ?project . }

  ?dmp dct:identifier ?dmpId ;
        dcso:identifierType ?dmpIdType .

  ?contact foaf:name ?author ;
           foaf:mbox ?email .
}
```

# Preparing maDMPs

- input data: Zenodo Community Data Stewardship 2021 – DMPs

[4] <https://zenodo.org/communities/dast-2021/>

- ensure conformity with JSON schema

- JSON-LD serialization (instances of DCSO)

[5] <https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/tree/master/ontologies>

```
"distribution": [
  {
    "title": "Raw data",
    "description": "Number of users of Facebook, Twitter and Instagram on quarterly basis from 2010 to 2020, missing values are represented as -",
    "format": ["text/csv"],
    "byte_size": 1154,
    "data_access": "open",
    "license": [
      {
        "license_ref": "https://www.statista.com/imprint/",
        "start_date": "2021-04-20"
      }
    ]
  }
]
```

[6] Winkler, M.: Machine-actionable DMP: Impact of social media on suicide rates. Zenodo (2021). <https://doi.org/10.5281/zenodo.4701948>

```
{
  "@id": "_:b7",
  "dataAccess": "open",
  "dcat:byteSize": 1154,
  "description": "Number of users of Facebook, Twitter and Instagram on quarterly basis from 2010 to 2020, missing values are represented as -",
  "format": "text/csv",
  "hasLicense": "_:b8",
  "title": "Raw data"
},
{
  "@id": "_:b8",
  "licenseRef": "https://www.statista.com/imprint/",
  "startDate": "2021-04-20"
},
```

# Use Case Application

- evaluate maDMPs using SPARQL queries
- completeness of maDMPs
- satisfaction value (SV): scale of 0 to 5

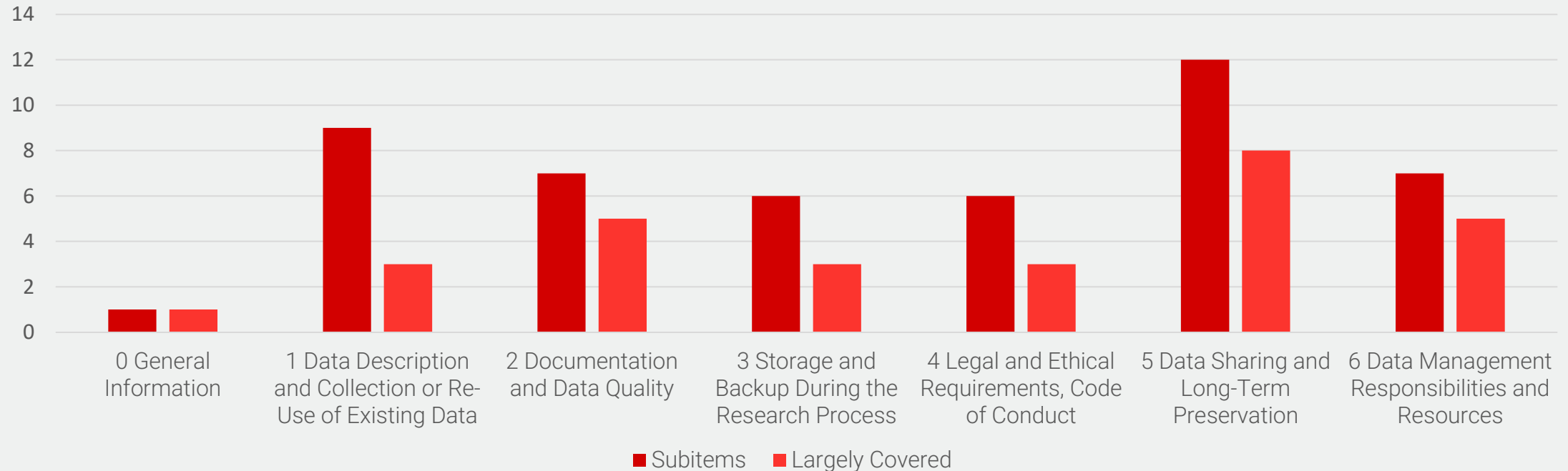
CATEGORY	SATISFACTION VALUE	JUSTIFICATION
0 General Information	2	Sufficient information about DMP. Information about project not included.
1 Data Description and Collection or Re-Use of Existing Data	2	The size of the produced/used data is provided. However, for two out of four distributions, the description is missing. Furthermore, the file formats of the produced data are not specified (in contrast to the reused data).
2 Documentation and Data Quality	2	No information about metadata or versioning provided. Keywords are included for half of the defined datasets. Minimal information about naming conventions included, as well as some statements about quality assurance measures.
3 Storage and Backup During the Research Process	2	maDMP does not have <i>host</i> elements defined, therefore some information is missing (backup type and frequency, availability). Good description of access restrictions. For most datasets, clear indication whether personal/sensitive data is stored provided.
4 Legal and Ethical Requirements, Code of Conduct	5	There is no information about potential preservation considerations. Regarding licenses, the maDMP does contain helpful data. However, the SPARQL query is a little bit too strict and fails due to the missing host definition. Good description of access restrictions and sufficient declaration of ethical considerations.
5 Data Sharing and Long-Term Preservation	3	maDMP does not have <i>host</i> elements defined, therefore a lot of important information is missing (PID system, backup strategies, URLs etc.). There are preservation statements in the original JSON file, but they cannot be queried from the JSON-LD due to the reason explained above. Regarding licenses (license, embargo, openness, sensitivity), the maDMP does contain helpful data. However, the SPARQL query is a little bit too strict and fails due to the missing host definition.
6 Data Management Responsibilities and Resources	1	Contact person is defined, but no contributors and their roles. Costs (resources, equipment, staff expenses etc.) are not specified in the maDMP.
<b>Sum</b>	<b>17/35</b>	

Due to the missing *host* definition, a lot of information could not be extracted with the queries. There is virtually no documentation of metadata. Information about the data management responsibilities is missing as well. Apart from those aspects, the maDMP provides a decent informational value.

# Mapping – Coverage

28/48 subitems largely covered (58%)

Coverage of Science Europe Evaluation Rubric Categories



detailed coverage report in GitHub repository

[7] Foidl, R., Brugger, L.: Evaluation of maDMPs using SPARQL (Jul 2021), <https://doi.org/10.5281/zenodo.4997671>

# Use Case Application

Category	Average SV
0 General Information	3.9
1 Data Description and Collection or Re-Use of Existing Data	4.0
2 Documentation and Data Quality	1.6
3 Storage and Backup During the Research Process	2.3
4 Legal and Ethical Requirements, Code of Conduct	3.2
5 Data Sharing and Long-Term Preservation	3.6
6 Data Management Responsibilities and Resources	3.5
Sum	22/35

complete evaluation results in GitHub repository

[7] Foidl, R., Brugger, L.: Evaluation of maDMPs using SPARQL (Jul 2021), <https://doi.org/10.5281/zenodo.4997671>

# Limitations

- some criteria not covered by maDMP schema
- necessary to make assumptions
- limited to information collection and filtering (no interpretation)
- cannot replace manual assessment

# Conclusion

- first step towards helping funders benefit from maDMPs
- filter relevant information and create custom views
- validate fulfillment of certain requirements
- SPARQL queries especially useful for
  - general information
  - data management responsibilities and resources
  - documentation and data quality

# Next Steps

- Shapes Constraint Language (SHACL) or Shape Expressions (ShEx) to model funder-specific requirements
- integrate tools that act upon information in maDMPs (e.g. F-UJI)
- consider funder-specific extensions of RDA DMP Common Standard

# Our Paper

## Automating Evaluation of Machine-Actionable Data Management Plans with Semantic Web Technologies

Raffael Foidl<sup>1</sup>[0000-0002-1748-1571], Lea Salome Brugger<sup>1</sup>[0000-0001-8770-4112],  
and Tomasz Miksa<sup>1,2</sup>[0000-0002-4929-7875]

<sup>1</sup> TU Wien, Vienna, Austria

<sup>2</sup> SBA Research, Vienna, Austria

**Abstract.** Machine-actionable data management plans (maDMPs) have, by their very nature, potential to bring advantages over data management plans that are written in text form. By employing maDMPs, not only researchers should be able to benefit from their merits, but also research funders receiving and assessing the DMPs. Science Europe, which is an association of major European research funders, have published an evaluation rubric that provides a common basis to support the evaluation of DMPs. By stating a set of criteria, it helps to ensure submitted DMPs cover required aspects and support FAIR data management. In this paper, we present a semi-automatic approach to leverage the benefits of maDMPs by providing SPARQL queries that represent requirements of Science Europe. The goal is to support reviewers in the assessment of DMPs expressed as maDMPs. The results show that semantic web technologies can help in providing customised views to reviewers, but human inspection and interpretation is still needed.

**Keywords:** Machine-Actionable Data Management Plan · Reviewing · Funder Support · RDF · JSON-LD · SPARQL

raffaelfoidl / maDMP-evaluation Public

<> Code Issues Pull requests Actions Projects Wiki Security Insights

main 2 branches 4 tags Go to file Add file Code

File/Folder	Commit Message	Commit Date
raffaelfoidl	Fix DOI in CITATION.cff	1753331 on Sep 29 69 commits
.github/workflows	Also allow empty commit during documentation page r...	6 months ago
app-bin	Change folder of packaged dcso-json tool from bin to ...	6 months ago
dcso-json	Fix broken links, typos and incorrect chapter references	3 months ago
docs	Generate documentation	last month
maDMPs	Change folder of packaged dcso-json tool from bin to ...	6 months ago
queries	Add SPARQL queries and adjust README for category 0	4 months ago
.gitignore	Change folder of packaged dcso-json tool from bin to ...	6 months ago
CITATION.cff	Fix DOI in CITATION.cff	last month
LICENSE	Initial commit	6 months ago
README.md	Fix broken links, typos and incorrect chapter references	3 months ago
generate_docs.sh	Add README of dcso-json into website posts, add "bac...	5 months ago

**About**

A project to facilitate the automatic assessment of the quality of machine-actionable data management. Especially conceived SPARQL queries to be executed over a DCSO serialization - in this instance, JSON-LD - allow for making solid assumptions about the informational value of a maDMP.

[raffaelfoidl.github.io/madmp-evaluation/](https://raffaelfoidl.github.io/madmp-evaluation/)

json sparql rdf assessment  
json-ld fair dmp fair-data  
reviewing madmp

Readme  
MIT License  
Cite this repository

[7] Foidl, R., Brugger, L.: Evaluation of maDMPs using SPARQL (Jul 2021), <https://doi.org/10.5281/zenodo.4997671>

[8] Foidl, R., Brugger, L., Miksa, T.: Automating Evaluation of Machine-Actionable Data Management Plans with Semantic Web Technologies. In DaMaLOS – Second Workshop on Data and Research Objects Management for Linked Open Science: Co-located at the International Semantic Web Conference ISWC 2021. PUBLISSO (Nov 2021), <https://doi.org/10.4126/FRL01-006429413>

# References

- [1] Science Europe: Practical Guide to the International Alignment of Research Data Management – Extended Edition (Jan 2021), <https://doi.org/10.5281/zenodo.4915862>
- [2] Miksa, T., Walk, P., Neish, P.: RDA DMP Common Standard for Machine-actionable Data Management Plans (2020). <https://doi.org/10.15497/rda00039>
- [3] Cardoso, J., Garcia, L., Ekaputra, F., Jacquemot, M., Miksa, T., Borbinha, J.: Towards semantic representation of machine-actionable Data Management Plans. . In DaMaLOS – First Workshop on Data and Research Objects Management for Linked Open Science: Co-located at the International Semantic Web Conference ISWC 2020. PUBLISSO (Nov 2020), <https://doi.org/10.4126/FRL01-006423289>
- [4] <https://zenodo.org/communities/dast-2021/>
- [5] <https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/tree/master/ontologies>
- [6] Winkler, M.: Machine-actionable DMP: Impact of social media on suicide rates. Zenodo (2021). <https://doi.org/10.5281/zenodo.4701948>
- [7] Foidl, R., Brugger, L.: Evaluation of maDMPs using SPARQL (Jul 2021), <https://doi.org/10.5281/zenodo.4997671>
- [8] Foidl, R., Brugger, L., Miksa, T.: Automating Evaluation of Machine-Actionable Data Management Plans with Semantic Web Technologies. In DaMaLOS – Second Workshop on Data and Research Objects Management for Linked Open Science: Co-located at the International Semantic Web Conference ISWC 2021. PUBLISSO (Nov 2021), <https://doi.org/10.4126/FRL01-006429413>

# Mapping – Coverage

Category	Subitems	Largely Covered	Percentage
0 General Information	1	1	100%
1 Data Description and Collection or Re-Use of Existing Data	9	3	33%
2 Documentation and Data Quality	7	5	71%
3 Storage and Backup During the Research Process	6	3	50%
4 Legal and Ethical Requirements, Code of Conduct	6	3	50%
5 Data Sharing and Long-Term Preservation	12	8	67%
6 Data Management Responsibilities and Resources	7	5	71%
Sum	48	28	58%