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

TU Wien, SBA Research

Vienna, Austria

Raffael Foidl

raffael.foidl@gmail.com

Lea Salome Brugger

leasalome.brugger@gmail.com

Tomasz Miksa

TMiksa@sba-research.org

Agenda

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

Data Management Plan (DMP)

- formal document
- awareness tool
- stakeholders:
  - researchers create DMP
  - reviewers assess DMP
  - funders provide guidelines for DMP
- Science Europe Practical Guide to the International Alignment of Research Data Management

Science Europe evaluation rubric

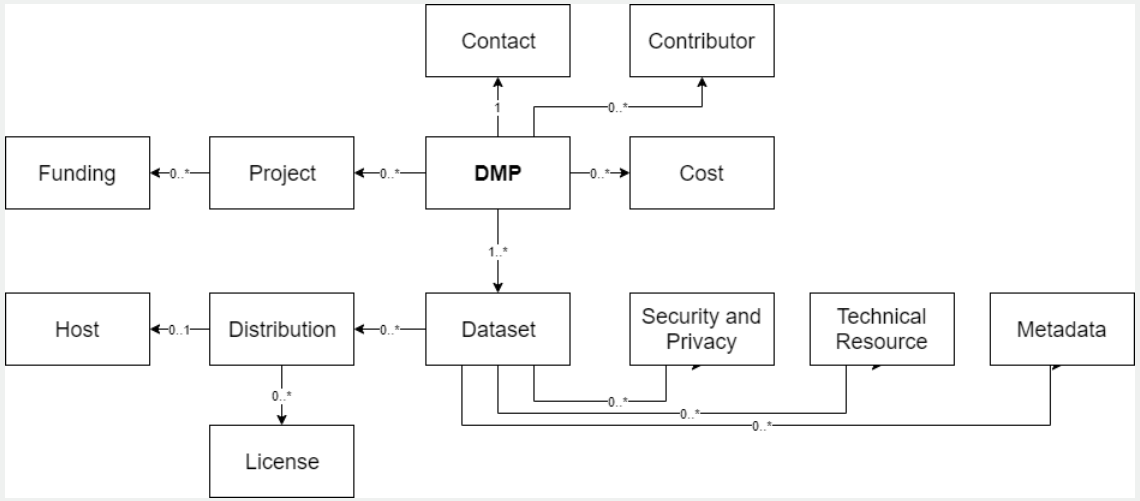
6 DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES			
Guidance for Researchers		**Sufficiently Addressed The DMP…**	**Insufficiently Addressed The DMP…**
6a **Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?**	• Outline the roles and responsibilities 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. • For collaborative projects, explain the co-ordination of data management responsibilities across partners • Indicate who is responsible for implementing the DMP, and for ensuring it is reviewed and, if necessary, revised. • Consider regular updates of the DMP.	• 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. • Clearly indicates who is responsible for day-to-day implementation and adjustments to the DMP. • Explains, for collaborative projects, the co-ordination of data management responsibilities across partners.	• 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. • Provides no description, in case of a collaborative project, on how data management responsibilities will be co-ordinated across partners.
6b **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)?**	• Explain how the necessary resources (for example time) to prepare the data for sharing/preservation (data curation) have been costed in. • Carefully consider and justify any resources needed to deliver the data. These may include storage costs, hardware, staff time, costs of preparing data for deposit, and repository charges. • Indicate whether additional resources will be needed to prepare data for deposit or to meet any charges from data repositories. If yes, explain how much is needed and how such costs will be covered.	• 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.	• 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.

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

Machine-Actionable Data Management Plan (maDMP)

- creation and assessment of DMPs time-consuming
- capture key information
- allow exchange of DMPs between systems
- RDA DMP Common Standard

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"]
                }
              },
```

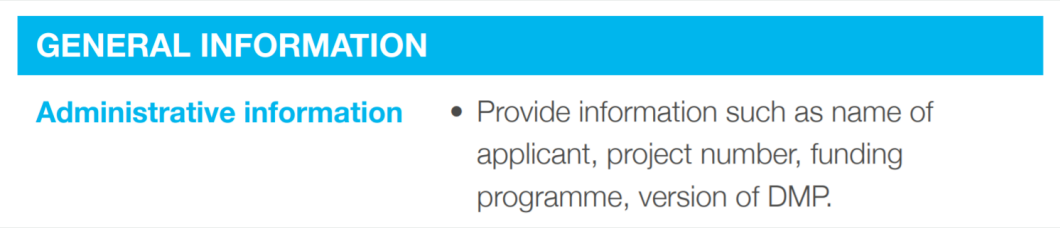

[2] Miksa, T., Walk, P., & Neish, P. (2019). RDA DMP Common Standard for Machine-actionable Data Management Plans. https://doi.org/10.15497/rda00039

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

Mapping

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



[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
  [3] https://zenodo.org/communities/dast-2021/
- ensure conformity with JSON schema
- JSON-LD serialization (instances of DCSO)
  [4] 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"
                        }
                ]
        }
]
```


[5] Winkler, Martin: 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

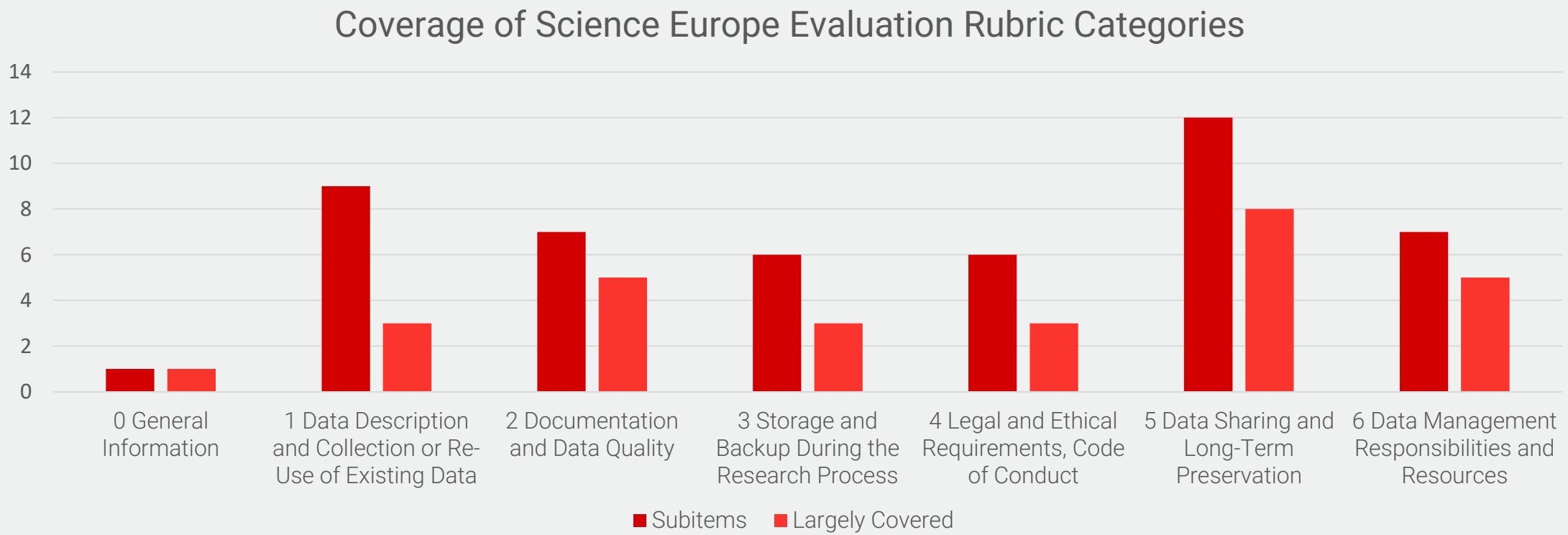
- evaluate maDMPs using SPARQL queries
- completeness of maDMPs
- satisfaction value (SV): scale of 0 to 5

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 `host` 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 `host` 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.
Sum	**17/35**	

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%)



detailed coverage report in GitHub repository

[6] 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

[6] 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

- 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
- future work
  - Shapes Constraint Language (SHACL), Shape Expressions (ShEX)
  - funder-specific extensions

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. (2019). RDA DMP Common Standard for Machine-actionable Data Management Plans. https://doi.org/10.15497/rda00039

[3] https://zenodo.org/communities/dast-2021/

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

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

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

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%